2016

Risk Classification in High Dimensional Survival Models

Daven Amin
University of Rhode Island, amind@my.uri.edu

RISK CLASSIFICATION IN HIGH DIMENSIONAL SURVIVAL MODELS

BY

DAVEN AMIN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2016

MASTER OF SCIENCE THESIS

OF

DAVEN AMIN

APPROVED:

Thesis Committee:

Major Professor Steffen Ventz

Gavino Puggioni

Corey Lang

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2016

ABSTRACT

Sparse regression models are an actively burgeoning area of statistical learning research. A subset of these models seek to separate out significant and non-trivial main effects from noise effects within the regression framework (yielding so-called "sparse" coefficient estimates, where many estimated effects are zero) by imposing penalty terms on a likelihood-based estimator. As this area of the field is relatively recent, many published techniques have not yet been investigated under a wide range of applications. Our goal is to fit several penalty-based estimators for the Cox semi-parametric survival model in the context of genomic covariates on breast cancer survival data where there are potentially many more covariates than observations. We use the elastic net family of estimators, special cases of which are the LASSO and ridge regression. Simultaneously, we aim to investigate whether the finer resolution of next-generation genetic sequencing techniques adds improved predictive power to the breast cancer patient survival models. Models are compared using estimates of concordance, namely the c-statistic and a variant which we refer to as Uno's C. We find that ridge regression models best fit our dataset. Concordance estimates suggest finer resolution genetic covariates improve model predictions, though further work with more observations is required.

ACKNOWLEDGMENTS

It's strange that the thesis, an attempt to boil down several months worth of effort, is an order of magnitude larger than the acknowledgements, which represent years of good advice, knowledge, love, and support given to the author. Perhaps that's because putting gratitude into words is very hard — harder than writing more thesis pages, anyway!

The Computer Science and Statistics department is a warm and vibrant place. I'm thankful for my interactions with the faculty and staff. Lorraine Berube and Beth Larimer are the pillars which keep the foundation from crumbling.

I am also thankful for

my thesis advisor, Dr. Steffen Ventz, who carefully shaped this work, kept me on track with gentle encouragement, and always provided timely feedback despite my tendency to work in fits and starts (usually aligned with looming deadlines),

the defense committee members, Dr. Gavino Puggioni and Dr. Corey Lang, whose courses reinforced my decision to jump into this field and who kindly agreed to help review this work,

my family, a fountain of unconditional love, who supported my return to full-time studies without hesitation, and

my friends, especially the new ones made at URI, who are the primary reason for any sanity I still retain.

The most important (and heartfelt) acknowledgement I can make is to Dr. Liliana Gonzalez. She has been a guardian angel at every stage of this adventure. She taught the first Statistics classes which kindled my desire to join the program. She offered me the opportunity to immerse in studies full-time on a graduate assistantship.

She and Dr. Puggioni nominated me as a graduate assistant to support URI's Coastal Resources Center, giving me the chance to spend two months in Accra, Ghana. She and her husband also took me in as a tenant while I finished up the program (weekly bicycle rides with John helped keep stress levels in check this past year!) I wouldn't be a member of the Statistics program, let alone writing a statistical work, if not for Liliana.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

Introduction and Background

Information is being collected at an ever-increasing rate, yet more information does not guarantee its holder more knowledge. The areas of statistical learning and machine learning have erupted with new techniques to automate the process of gleaning knowledge from massive stores of information. This thesis is an exploration of a sliver of these methods, applied to a specific problem domain: cancer progress prediction through genomic sequencing. We investigate whether finer-grained genomic information, available through newer sequencing methods, provides additional predictive power over the previous resolution of information.

1.1 Gene Expression

The central dogma of molecular biology describes the mechanisms by which information encoded at a cellular level is expressed as proteins by an organism (Crick et al., 1970). In a simplified form, it suggests information, which partially determines the physical traits of the organism, is encoded as linear sequences of deoxyribonucleic acid (DNA). This information is used to construct proteins, large three dimensional structures which facilitate nearly all cellular functions within the organism. A gene is a sequence of DNA that maps to a particular protein, and the genome comprises all genes. Genes are expressed as proteins through a series of steps. DNA is transcribed into an intermediate ribonucleic acid (RNA) form, and the RNA sequence is spliced to remove non-coding "intron" regions and join the remaining protein-coding "exon" regions into new contiguous sequences (Gilbert, 1978). The order and sequence of the exon regions determines the specific protein variant which will be generated. Multiple splicing variants of a gene may exist. The splicing variants of a gene, referred to as its "isoforms", use different arrangements and subsets of ex-

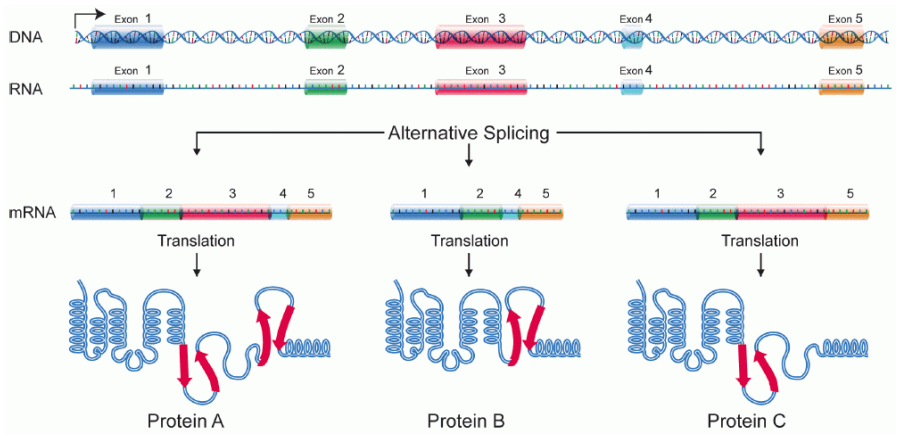

Figure 1. Genes, encoded as DNA, are expressed as proteins through a multi-stage process beginning with transcription into RNA. Protein coding exon regions are then spliced and translated into protein structures. Gene expression can be measured by counting mRNA splicing variants. The expression can either be treated as the sum or the set of all mRNA counts derived from a gene. Source: National Human Genome Research Institute

ons. Spliced RNA sequences are translated directly into protein structures by cellular mechanisms. Gene expression can be thought of as either the sum of all proteins derived from a single gene's DNA or specific levels of the protein variants derived from a gene's set of isoforms.

Many diseases, such as breast cancer, arise from corruption of the genetic code within exon regions (the exome) (Bishop, 1987). Cancer is the manifestation of uncontrolled cell growth. Since cellular functions are dependent on protein, excessive upregulation of protein production can increase cell reproduction. Similarly, certain regulatory functions are maintained through protein structures. In these cases, downregulation of these proteins could also result in increased cell reproduction. The cell constantly dismantles and regenerates many proteins in order to facilitate these regulation mechanisms. Measuring the level of gene expression within a cell is an indicator of protein activity. This thesis aims to compare isoform-level expression information with overall gene expression in cancerous cells.

1.2 Survival Models

A survival model, or time-to-event model, is a set of assumptions made about the survivor function of a random event. The survivor function, denoted $S(t)$, is defined as the probability of the random event occurring after some time t, denoted $\Pr(T > t)$ where T is the random variable *time of event* and both t and T are defined in relation to a common starting time. The survivor function is the complement of the cumulative distribution function of T (Kleinbaum and Klein, 1996). We assume the event will eventually occur such that T is strictly positive and continuous. Sometimes T is referred to as the "failure time". If we assume $S(t)$:

- equals 1 when $t = 0$ (the event cannot occur until some time has passed),
- is non-increasing for $t > 0$, and
- approaches 0 as $t \to \infty$ (the event must occur given infinite time).

then we can show $S(t)$ has a one-to-one relationship with a hazard function $h(t)$ which represents the instantaneous rate of the event occurring at time t given that it has not yet occurred. The relationship between survivor function and hazard function is given by

$$h(t) = -[\frac{dS(t)/dt}{S(t)}]$$
$$S(t) = e^{-\int_0^t h(u)du}.$$

The hazard at time t is a rate, not a probability, and has less restrictions on its form. Namely, a valid hazard function yields non-negative values for all values of t and integrates to infinity over the positive real line. By modeling a valid hazard function we've effectively modeled a valid survivor function using the relationships shown above.

Modeling the hazard function allows us to answer questions in a similar fashion to modeling the log-odds of the event occurring, such as the determination of risk

factors in a logistic regression. However, the hazard function explicitly accounts for the passage of time. This can yield more information from "right-censored" observations when fitting a model. Right-censored observations are when the event must have occurred after a time t_i, but we do not know the actual time-of-event. Modeling the log-odds ignores the censoring time t_i and only considers the absence of the event. Modeling the hazard rate, however, uses both censoring times and time of events during the model fitting, as discussed below in the specific case of the Cox proportional hazards model.

1.2.1 The Cox Proportional Hazards Model

One way to model the effect of covariates on the time of a random event uses the approach of hazard regression. If the covariates X are unchanging with respect to time, their effect on the hazard function can be modeled as $h(t|X) = h(t)g(X)$ where $g(X) > 0$. A popular model family for this conditional hazard is based on the Cox semi-parametric proportional hazards model (Cox, 1972). The Cox model assumes that the ratio of two rates with different covariates, i.e. conditional hazard ratio, for the same event is constant and a function of the covariates. The covariates modify the shared underlying hazard function $h_0(t)$, denoted as the baseline hazard.

$$\frac{h(t|X_i)}{h(t|X_j)} = e^{(X_i - X_j)\beta}$$

$$h(t|X_i) = h_0(t) \times e^{X_i\beta}$$

$$\log h(t|X_i) = \gamma(t) + X_i\beta$$

$$S(t|X_i) = S(t|X_j)^{\exp((X_i - X_j)\beta)}$$

The log hazard equation above shows how the Cox model may be seen as a generalized linear model (Nelder and Baker, 1972), where the intercept $\gamma(t)$ contains the baseline hazard and a log link is used to estimate the hazard rate rather than an expected value. Extensions to the Cox model allow for time-varying covariate ef-

fects (Fisher and Lin, 1999) but lose the ability to predict individual survivor functions. In this thesis, however, we do not consider any time-varying covariates.

The Cox model is semi-parametric as it assumes linear covariate effects ($\sum X_i\beta$) on the hazard rate and it fits the effects β (also referred to as the coefficients of the model) to observations using maximum likelihood of a partial likelihood function which does not specify the baseline hazard. A fully parametric model with a traditional likelihood function would have to specify, and thus make assumptions about, the baseline hazard. The partial likelihood function, $\mathbb{L}_{PL}(\beta)$, for the Cox model is given by

$$\mathbb{L}_{PL}(\beta) = \prod_{r\in D} \frac{\exp(X_{j_r}\beta)}{\sum_{j\in R_r}\exp(X_j\beta)}$$

where D is the set of event times, R_r is the set of observations at risk of the event at time r, and j_r is the observation with the event at time r (Tibshirani et al., 1997). Maximum partial likelihood suggests the values of β which produce the highest value of the partial likelihood function given a dataset are the best estimates for the true parameters of the model. In general, maximum partial likelihood estimation is solved as an optimization problem. For mathematical and computational convenience, this is often reformulated as maximizing the log partial likelihood function $\log\mathbb{L}_{PL}(\beta)$, and the maximum partial likelihood estimator (MPLE) of β, $\hat{\beta}$, is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\{\log\mathbb{L}_{PL}(\beta)\}.$$

Fitting a Cox model allows a researcher to make probabilistic statements about relative risk (i.e a patient who smokes has twice the risk of experiencing a heart attack as one who does not smoke, all else being equal.) If the covariates are continuous, e^{β_j} can be considered a the change in relative risk due to a one unit change of covariate X_j.

1.3 Regularized Models

In problems where there are more unknown parameters to be fitted than observations, the Cox model is over-parameterized; that is, more than one combination of parameter estimates maximize the partial likelihood function. Some other information would be needed to distinguish the "best" set of parameter estimates. Modifying this over-parameterized model such that the MPLE yields a single unique estimate is termed "regularizing" the model. If a regularized likelihood function is convex, the MPLE of the regularized model may be estimated using efficient convex optimization techniques. This is important to fit models on large datasets in a reasonable amount of time (i.e. hours or days!)

Moreover, if it is known apriori that not all, or possibly just a few, covariates affect the time of the outcome event then the model should also account for this. The assumption of relatively few true covariate effects is termed "sparsity" (James et al., 2013; Hastie et al., 2015). This thesis considers several cases of one family of regularization methods which exploit the sparsity assumption. In these methods β is estimated using a "penalized" estimator which can be written as

$$\hat{\beta}^* = \underset{\beta}{\operatorname{argmax}}\{\log \mathbb{L}_{PL}(\beta) - \lambda PT(\beta)\}.$$

λ is a free parameter which dictates the bias towards 0 exerted by a penalty term $PT(\beta)$, where $PT(\beta) \geq 0$. If $\lambda = 0$ and the model is not over-parameterized, $\hat{\beta}^*$ yields the maximum likelihood estimate given by $\hat{\beta}$. As λ approaches ∞, the estimator $\hat{\beta}^*$ becomes smaller and smaller, with all covariate effects becoming 0. A penalized estimator will give estimates which are biased towards 0. In the context of prediction this property may be advantageous even if the underlying model is not over-parameterized. If the data contain many "noisy" measurements or covariates which do not truly influence the outcome, reducing the magnitudes of the estimated covariate effects may reduce model overfitting.

1.3.1 Ridge Regression, LASSO, and Elastic Net

The ridge (Hoerl and Kennard, 1970) estimator is a penalized estimator which uses a penalty term of the squared Euclidean distance, or L_2^2 norm, of the coefficients. The L_2^2 norm of β is defined as

$$\|\beta\|_2^2 = \sum_j {\beta_j}^2.$$

This norm shrinks coefficients towards zero as λ increases, but the estimator does not produce sparse estimates. Rather, all coefficients are reduced in magnitude. Hoerl et al. originally defined the ridge estimator in the context of multiple linear regression, where they proved the ridge estimator always has a value of λ for which it produces β estimates with lower mean squared error than the regular Ordinary Least Squares estimator. Further work by Cessie and others (Le Cessie and Van Houwelingen, 1992) showed this is asymptotically true for a ridge estimator in the case of logistic regression, and that a ridge estimator can be applied to the Cox model (Verweij and Van Houwelingen, 1994). Additionally, the penalty function used to find the ridge estimator is convex and so the ridge estimator is unique and may be found quickly even for large datasets.

A related penalized estimator is the LASSO (Tibshirani, 1996), or lasso, estimator. LASSO is short for Least Absolute Shrinkage and Selection Operator. The lasso estimator uses a penalty term of the Manhattan distance, or the L_1 norm, of the coefficients. The L_1 norm of β is defined as

$$\|\beta\|_1 = \sum_j |\beta_j|.$$

The L_1 norm has several useful properties which have boosted the popularity of the lasso. It is the "smallest" of the L_p family of norms (in terms of p) which is convex, and so it yields an estimator with a convex penalty function (Hastie et al., 2015). The "smallest" member of the L_p family of penalty terms is the L_0 generalized norm which is optimal for variable selection, as it corresponds to best subset selection where the

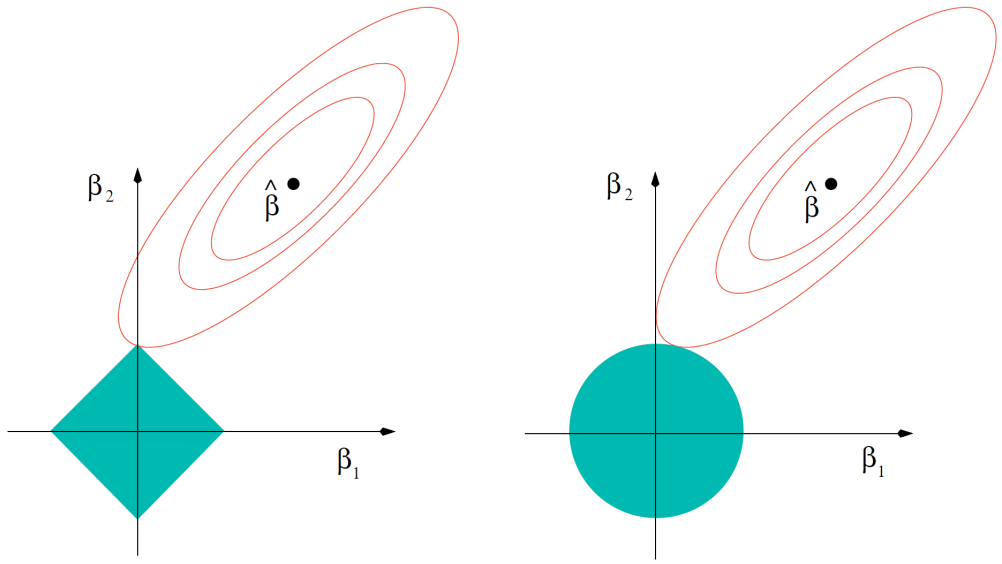

Figure 2. The lasso tends to produce sparse coefficient estimates, while the ridge estimator shrinks all estimates towards zero. In the case of estimating two coefficients, the free parameter λ in the lasso estimator dictates the size of a square region (left) corresponding to the L_1 norm penalty, while the ridge estimator's L_2^2 penalty creates a circular constraint (right). The estimators return the values in the parameter space closest to the maximum likelihood estimate ($\hat{\beta}$) which meet their constraints. In the case of the lasso, this tends to be on a vertex of the square (or edge of the hypercube in higher dimensions), yielding one or more coefficient estimates of zero. Source: Statistical Learning with Sparsity

likelihoods of all combinations of inclusions for predictor variables are considered. Best subset selection, however, requires computational resources which are infeasible for non-trivial problems. The L_1 norm as a penalty term tends to remove predictor variables from fitted models, allowing the lasso estimator to yield "sparse" estimates. The use of the L_1 penalty also has the propensity to shrink small-magnitude covariate effects to 0 while leaving large-magnitude effects close to their unbiased MPLE estimates. However, if the covariates are highly correlated the lasso estimator will arbitrarily shrink one effect towards zero.

The elastic net (Zou and Hastie, 2005) estimator attempts to yield sparse estimates similar to the lasso while imposing uniform shrinkage on the effects of covariates by using a weighted combination of the L_1 and L_2^2 norms as the penalty term.

It introduces a second free parameter, $\alpha \in (0,1)$, and uses the weighted combination $(\alpha \left\|\beta\right\|_1 + \frac{1}{2}(1-\alpha)\left\|\beta\right\|_2^2)$ as the penalty term. It can be thought of as a generalization of the ridge and lasso estimators.

Bayesian Interpretation

These estimators can also be considered in the context of the Bayesian framework. The Bayesian view of modeling focuses on obtaining the posterior distribution of a model's coefficients $\Pr(\beta|D)$, given an observed dataset $D = (Y, C, X)$ (composed of time-to-events Y, censoring indicators C, and covariates X) and a prior (possibly uninformative) distribution of the model coefficients $\Pr(\beta)$. This posterior distribution is proportional to model likelihood multiplied by the prior distribution.

$$\Pr(\beta|D) \propto \Pr(D|\beta)\Pr(\beta)$$

$$\Pr(\beta|D) \propto \mathbb{L}(\beta)\Pr(\beta)$$

The "maximum a-posteriori" or MAP estimate of β can be obtained as the mode of $\Pr(\beta|D)$ and is equivalent to maximizing $\log\Pr(\beta|D)$, yielding the equation

$$MAP_\beta = \underset{\beta}{\operatorname{argmax}}\{\log\mathbb{L}(\beta) + \log\Pr(\beta)\}.$$

Using the partial likelihood for the Cox model, the ridge estimator is equivalent to the MAP estimator where the effects β_i have i.i.d Gaussian prior distributions such that

$$\beta_i \overset{\text{i.i.d}}{\sim} N(0, \frac{1}{2\lambda})$$

(Hastie et al., 2015). The lasso estimator is equivalent to a MAP estimator using i.i.d Laplacian prior distributions on each β_i, centered at zero and with spread parameters determined by λ. The elastic net is equivalent to a MAP estimator using a mixture of penalty terms as the i.i.d priors on each β_i.

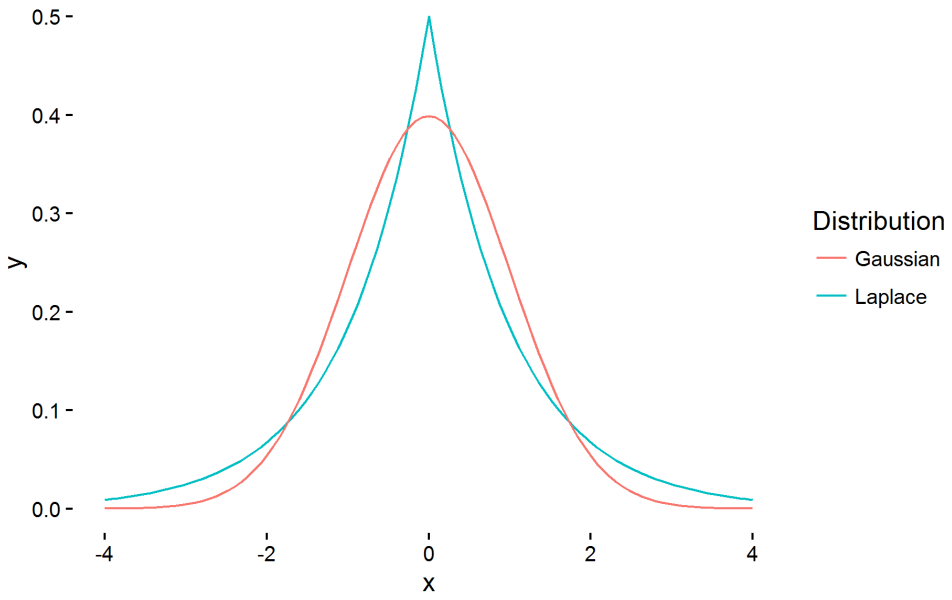

Figure 3. Comparison of Laplace and Gaussian distributions with identical location and scale parameters (0 and 1, respectively). The Laplace distribution places more probability mass at its central value and in its tails than the Gaussian distribution. This characteristic is indicative of the lasso estimator's sparsity properties.

1.4 Model Assessment

Many techniques to order statistical models from "better" to "worse" quantify the "fit" of each model. That is, they use criteria derived from each model's likelihood function. A well fitting model can then be interpreted in the context of its assumptions.

In contrast, we are interested in comparing the predictive power of our estimated models. Several measures have appeared in the literature, such as the cross-validated partial likelihood, Brier Score, and pseudo R^2 (Van Wieringen et al., 2009). Cross-validated partial likelihood is based on the model likelihood function, discussed previously and is presented in our results. Here we discuss another measure of predictive power for a single model (concordance), one of its interpretations (area under a receiver operating characteristics curve), and a variant of a common estimator for it. Due to the asymptotic normality of the estimator we can test for differences in pre-

dictive power between models, which we also present in our results.

1.4.1 Concordance and Receiver Operating Characteristics Curves

Concordance is defined as

$$\text{concordance} = P(\text{pred}(X_1) > \text{pred}(X_2) | T_1 < T_2)$$

where (T_i, X_i), $i = 1, 2$ are two independent failure times and covariate sets and $\text{pred}(X_i)$ is a predictor of the hazard rate as a function of the covariates. The estimated linear predictor, $\hat{\beta} X_i$, is used as $\text{pred}(X_i)$ in the case of Cox and penalized Cox models. In the absence of right-censoring the concordance is equivalent to the area under a Receiver Operating Characteristics (ROC) curve for the given model (Hanley and McNeil, 1982) and is related to the sensitivity and specificity of the model.

The ROC curve for a predictive model with two outcomes is a graphical representation of the model's ability to correctly predict observations experiencing the event or "positive" observations, referred to as sensitivity, as the tolerance for incorrect positive predictions or "false positives" ranges from nil to unbounded.

In a binary outcome scenario, where outcomes are "true" or "false" and we have a set of covariates for each "true" or "false" observation, a model which computes a score based on the covariates can be considered to predict a "true" outcome if the score is above a fixed threshold and "false" otherwise. The model's performance can be evaluated according to its sensitivity and specificity. The sensitivity is the proportion of true positives (the outcome was "true" and the model predicted "true") among all "true" observations, and the specificity is the proportion of true negatives (the outcome was "false" and the model predicted "false") among all "false" observations. As seen in Table 1, a confusion matrix showing the expected distribution of model predictions against actual outcomes can be constructed from the model's sensitivity and specificity along with a set of outcomes.

	Model: “true”	Model: “false”	Total
Actual: “true”	sensitivity × P	(1-sensitivity) × P	P
Actual: “false”	(1-specificity) × N	specificity × N	N

Table 1. A confusion matrix giving the expected distribution of a model’s prediction outcomes in terms of its sensitivity and specificity, given some P “true” outcomes and N “false” outcomes.

The ROC curve plots specificity against sensitivity as the threshold for class prediction is varied. The horizontal axis denotes (1-specificity) and the vertical axis denotes (sensitivity).

A model which has no predictive power will appear as a straight line between 0,0 (denoting the threshold at which all observations are classified as “false”), and 1,1 (denoting the threshold at which all observations are classified as “true”) since for any given threshold the proportion of true positives (sensitivity) is equal to the proportion of false positives (1-specificity). The area under this curve (AUC) is .5. The AUC of a model which always correctly classifies “true” observations is 1, since the proportion of true positives (sensitivity) equals 1 regardless of the proportion of false positives (1-specificity). If a model has an AUC less than .5 (a misclassifier) it can be converted to one with an AUC greater than .5 by swapping the predicted classification of observations.

There are several methods to assess whether two models fitted on the same observations have statistically different AUCs including those based on the asymptotic normality of concordance (DeLong et al., 1988).

1.4.2 C-Statistics

Survival models model time-to-event, not binary outcome, and must use adapted ROC methods. Instead of computing the full ROC curve to determine the AUC, a prevalent estimator of concordance for a survival model is the c-statistic (Lee and Mark, 1996). The c-statistic is computed by first taking all pairs of

observations where at least observation is not right-censored and assigning each pair a class label. The label is

- "concordant" if the observation with a smaller failure time experienced the event and had a higher predicted hazard than the other,
- "discordant" if the observation with a smaller failure time had a lower predicted hazard, or
- "tied" if the predicted hazards are equal and only one observation of the pair experienced the event.

Pairs where the observation with the smaller failure time did not experience the event are discarded, as are pairs where both observations experienced the event at the same time.

The c-statistic is defined as

$$\text{c} - \text{statistic} = \frac{\text{concord} + .5(\text{tied})}{\text{concord} + \text{discord} + \text{tied}}$$

where concord is the number of concordant pairs, discord is the number of discordant pairs, and tied is the number of tied pairs.

If all observations experienced the event, the c-statistic is an unbiased estimator of concordance. In the presence of right-censoring the distribution of the c-statistic is dependent on the censoring distribution and may no longer directly estimate the concordance (Koziol and Jia, 2009).

We note that detecting whether a survival model has increased predictive power over another may not be possible by the comparing the concordances if both models have reasonable predictive power (Pencina et al., 2008).

However, the use of concordance is still prevalent in the literature without a widely accepted alternative, though work is ongoing to introduce new measures

of predictive power such as variants of the Integrated Discrimination Improvement (Uno et al., 2013). We use the variant of the c-statistic from (Uno et al., 2011) to remove the effect of the censoring distribution on the estimate of concordance and test for differences in predictive discrimination between models.

List of References

Bishop, J. M. (1987). The molecular genetics of cancer. *Science*, 235(4786):305–311.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

Crick, F. et al. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157.

Gilbert, W. (1978). Why genes in pieces? *Nature*, 271(5645):501.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

Kleinbaum, D. G. and Klein, M. (1996). *Survival analysis*. Springer.

Koziol, J. A. and Jia, Z. (2009). The concordance index c and the mann–whitney parameter pr (x> y) with randomly censored data. *Biometrical Journal*, 51(3):467–474.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied statistics*, pages 191–201.

Lee, K. L. and Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387.

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*.

Pencina, M. J., D'Agostino, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. et al. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.

Uno, H., Tian, L., Cai, T., Kohane, I. S., and Wei, L. (2013). A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in medicine*, 32(14):2430–2442.

Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L. (2009). Survival prediction using gene expression data: a review and comparison. *Computational statistics & data analysis*, 53(5):1590–1603.

Verweij, P. J. and Van Houwelingen, H. C. (1994). Penalized likelihood in cox regression. *Statistics in medicine*, 13(23-24):2427–2436.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

CHAPTER 2

Methodology

Our dataset is part of data generated by The Cancer Genome Atlas Research Network (TCGA), a collaborative project between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). TCGA provides anonymized genomic and clinical data for major types and subtypes of cancer and makes the data freely available for research.

Specifically, we are focused on patient mortality of female breast cancer patients. We obtained two slices of TCGA's BRCA dataset from the Firehose portal of the Broad Institute (Broad Institute TCGA Genome Data Analysis Center, 2016), one containing clinical covariates and survival outcome of 1097 breast cancer patients and the other containing RSEM (Li and Dewey, 2011) estimated counts of gene isoforms for 1212 tissue samples.

2.1 Cleaning

The Firehose portal provides aggregated outputs from a pipeline sequence of tools. However, the raw output was not in a format immediately ready for analysis. Specifically, the data was organized in row-column format such that each row corresponded to a covariate (cancer stage, patient age, gene expression count) and each column corresponded to a patient. We used a set of R (R Development Core Team,) and Python (VanRossum and Drake, 2010) helper routines to transpose the raw data into a row-column format where each row represents a patient (an observation) and each column represents a variable of interest. Separate cleaning and transformation steps were required by each dataset (gene expression counts and survival time & clinical covariates) before merging the two for analysis in R as detailed below.

2.1.1 RSEM Data

In order to give researchers the ability to explore many aspects of breast cancer gene expression, the BRCA data contains gene expression estimates of tissue samples from breast tumors and normal non-cancerous regions. Though the underlying gene expression counts are integers, the RSEM count estimates take on real values. We removed the 112 normal tissue samples from the data, leaving 1100 sets of breast tumor gene expression estimates, with each set containing 73,599 isoform count estimates.

Equivalence of Gene and Isoform Counts

Although it was not used in our analysis, TCGA provides other datasets, including gene-level expression estimates. We aggregated the isoform expression dataset at the gene-level and compared it to the provided gene-level estimates to verify our aggregation procedure was correct. All (patient,gene-level) counts were within .01 units, which we attribute to rounding precision. We used this as justification for working directly with the isoform expression dataset and later aggregating it at the gene-level. 6229 isoforms out of the 73,599 total did not correspond to a known gene. We verified this was an anomaly of the underlying dataset and not an artifact of our procedures or an error. Although outside of the scope of this thesis, these isoforms do not map to a known gene in the HG19 reference genome used by the RSEM procedure to create expression count estimates. These isoforms were included in the normalization procedures detailed below if they exhibited non-trivial variance among patients, but are excluded from the analysis as we cannot compare them to a gene-level aggregate.

Removal of low-variance Isoform Counts

Many of the isoform expression estimates were zero for all patients. Since we are only interested in isoforms which have an effect on survival time, we removed

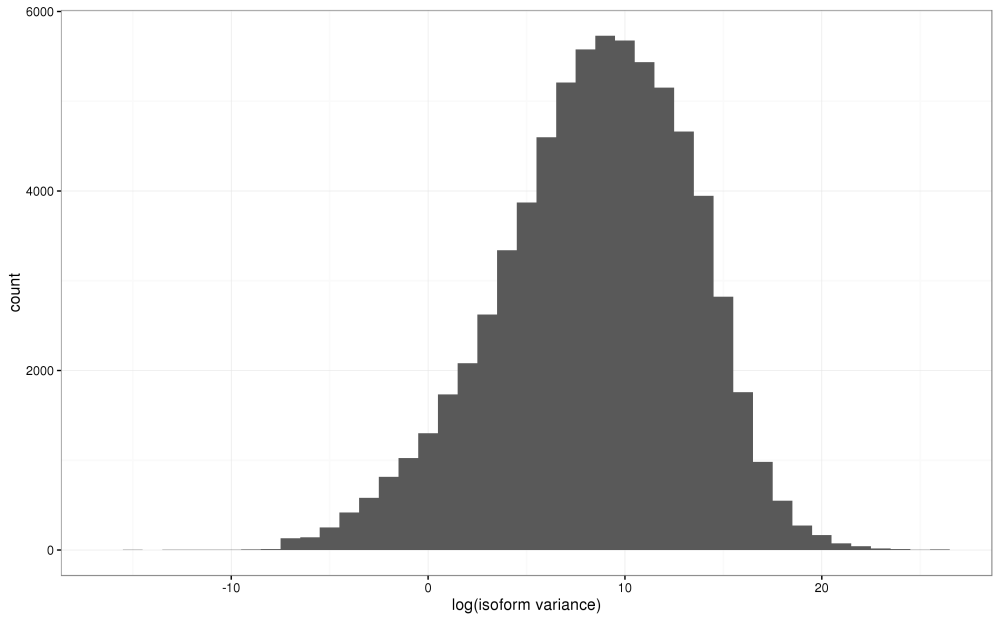

Figure 4. Histogram of log sample variance for estimated isoform counts in the original TCGA dataset. Isoforms with log sample variance less than or equal to zero were removed from the data.

all isoforms with low sample variance of expression counts among patients. Figure 4 shows the histogram of log sample variance for each isoform across all patients. Isoforms with log sample variance less than or equal to 0 were excluded from all subsequent cleanup and analysis, and are ignored for the rest of this thesis. 67,027 isoforms had sample variances of at least 1 and were not excluded, and 63,214 isoforms had a known gene correspondence. Although the 67,027 isoforms were used in the normalization procedures, our final RSEM data contained information from 63,213 isoforms corresponding to 19,330 genes.

Between and Within Sample Normalization

The isoform expression estimates are dependent on the amount of tissue sampled (with more biomass we would expect to see higher estimated counts) and need to be rescaled. A common scaled unit which we adopt is “counts per million” or CPM, which is just a scaled sample proportion of each estimated count per gene or isoform

Gene	Patient 1 Count	Patient 2 Count	Patient 1 CPM	Patient 2 CPM
1	2	6	.2	.06
2	2	6	.2	.06
3	2	6	.2	.06
4	2	6	.2	.06
5	2	76	.2	.76

Table 2. An example of within-sample normalization using counts per million (CPM) yielding misleading results: it is more likely that genes 1–4 are expressed identically in patients 1 and 2, while only gene 5 is differentially expressed. CPM normalization only considers the number of counts within each patient and yields a result suggesting all five genes are differentially expressed. Source: Harold Pimentel

i in patient j. In this section we will use the term "gene" to refer to both gene-level aggregates and isoforms when describing normalization techniques.

CPM, defined as

$$\text{CPM}_{i,j} = \frac{\text{count}_{i,j}}{\sum_i \text{count}_j} \times 10^6$$

only normalizes counts relative to a patient, or within-sample. An example from (Pimentel, 2014) illustrates the issue with this well; if we have counts from 5 genes for two patients where genes 1–4 are known to be expressed identically but gene 5 is expressed several orders of magnitude more in patient 2, we end up with the misleading CPMs shown in Table 2.

We need to normalize counts between patients as well in order to compare them. This requires the use of a between-sample normalization method. One method by (Bullard et al., 2010) uses an upper quantile based method to correct for between-sample differences. Normalized count data is available from Firehose using this method. We use the normalization method and implementation described by (Li et al., 2011) in which the "sequencing depth" of each patient is estimated using a log-linear model. They define sequencing depth as a measure of the relative counts between patients. In the example from Table 2 the sequencing depth of patient 2 would be 3 relative to patient 1, since genes 1–4 had raw counts three times higher

but are expected to show identical expression in both patients.

The normalization method is iterative, and uses half of the measured genes to estimate sequencing depth under the assumption that gene expression counts come from a Poisson distribution where the mean $\mu_{i,j}$ count for patient i and gene j is equal to the expression level of j scaled by the sequencing depth of i. A simplified description and example of the method in action is described below.

- In the initial step, the sequencing depth is estimated as the proportion of each patient's total gene counts. For the example in Table 2, the proportion vector would be approximately $(0.09, 0.91)$, by taking the marginal counts of $(10, 100)$ and dividing by the sum.

- Next, each total gene count is scaled by the current estimation of the sequencing depth. For Table 2, this would yield

$$\begin{array}{cc} 0.73 & 7.27 \\ 0.73 & 7.27 \\ 0.73 & 7.27 \\ 0.73 & 7.27 \\ 7.10 & 70.90 \end{array}$$

  These are the "expected" gene counts given the sequencing depth and the marginal gene counts.

- A goodness-of-fit (GOF) metric is calculated for each gene by summing the squared difference of the observed count and the expected count from the previous step divided by the expected count per gene patient pair.

$$\text{GOF} = \sum_{\text{gene}} \frac{(\text{observedcount} - \text{expectedcount})^2}{\text{expectedcount}}$$

  For the example, we have a GOF vector of approximately $(2.23 + 0.22, 2.23 + 0.22, 2.23 + 0.22, 2.23 + 0.22, 3.66 + 0.37)$ or $(2.45, 2.45, 2.45, 2.45, 4.02)$.

- The genes with GOF values within the first and third quartiles of the GOF vector are used to estimate the sequencing depth as in the initial step, and the procedure repeats. In the example, genes 1–4 would be used to estimate sequencing depth in the next iteration, and gene 5 would be ignored. When the estimates of sequencing depth remain nearly constant between iterations, the procedure can be terminated.

Our estimates of sequencing depth are summarized in Figure 5 and Table 3. The normalization procedure centers the final estimates around 1. The implementation also pre-filters genes with overall small counts before starting the procedure – this yielded 3012 isoforms out of 67,027 being filtered. We scaled each patient's isoform counts by the estimated sequencing depth to perform between-sample normalization. We then added 1 to each normalized count before converting each patient's normalized counts to CPM, guaranteeing all CPM values are greater than zero and allowing us to perform logarithmic transformations during the analysis.

Min. :0.2779	Max. :2.2771
Mean :1.0396	Median :1.0290
1st Qu.:0.8360	3rd Qu.:1.2304

Table 3. Descriptive statistics of estimated sequencing depth distribution. $n = 1100$.

2.1.2 Survival Times and Censoring

We used the "clinical pick" merged dataset, which contains high-level clinical covariates and survival outcomes for 1097 patients. We created the dependent variable *time* (time of event or censoring time) by using the "days_to_death" field when available or "days_to_last_followup" if missing. We also created a censoring indicator variable *censoring* with a value of 1 if days_to_death was used or 0 if days_to_last_followup was used.

The dataset contained observations from 12 males with breast cancer, of which

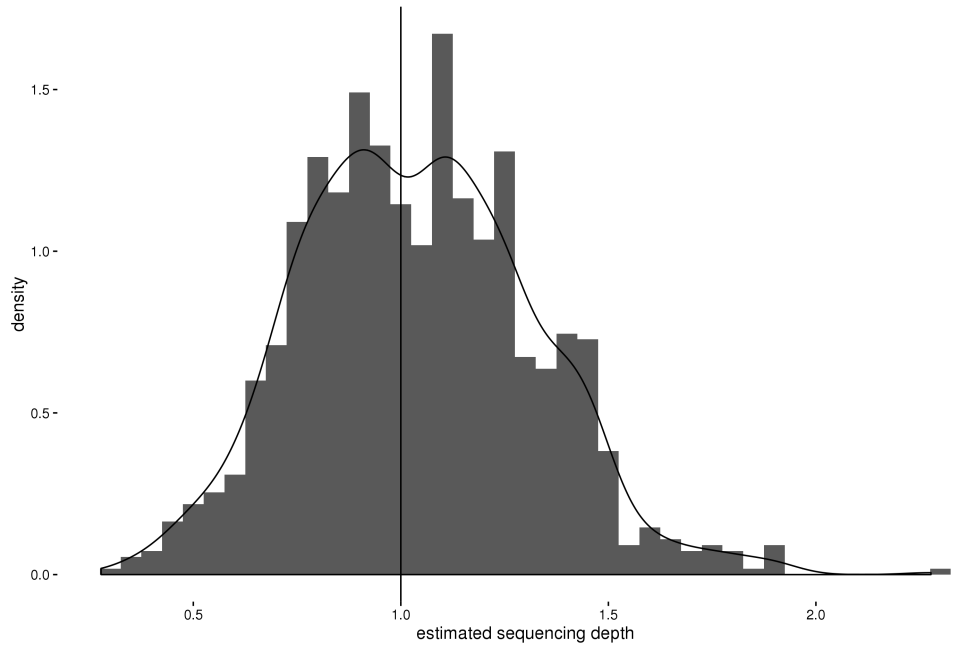

Figure 5. Relative histogram and smoothed density estimate of estimated sequencing depth distribution. Histogram counts have been scaled such that the total shaded area is equal to 1.

11 were censored observations. We removed these observations, yielding 1085 observations of female breast cancer patients, where 150 patients suffered mortality during their observation period and 14 were missing both a time of event and a censoring time. The median survival time was 3941 days. The Kaplan-Meier survival curve, along with 95% confidence intervals, is shown in Figure 6.

2.1.3 Merged Data and Clinical Covariates

We merged the isoform CPM transformed data and filtered "clinical pick" data using the common patient identifiers. This yielded a merged dataset containing 1074 observations. These observations contained all of the high-level clinical covariates from the "clinical pick" dataset.

We considered including several clinical covariates in our analysis as controlling variables. Originally we considered including "years_to_birth", "radiation_therapy", "number_of_lymph_nodes", "pathologic_stage", and "race". Age and number of

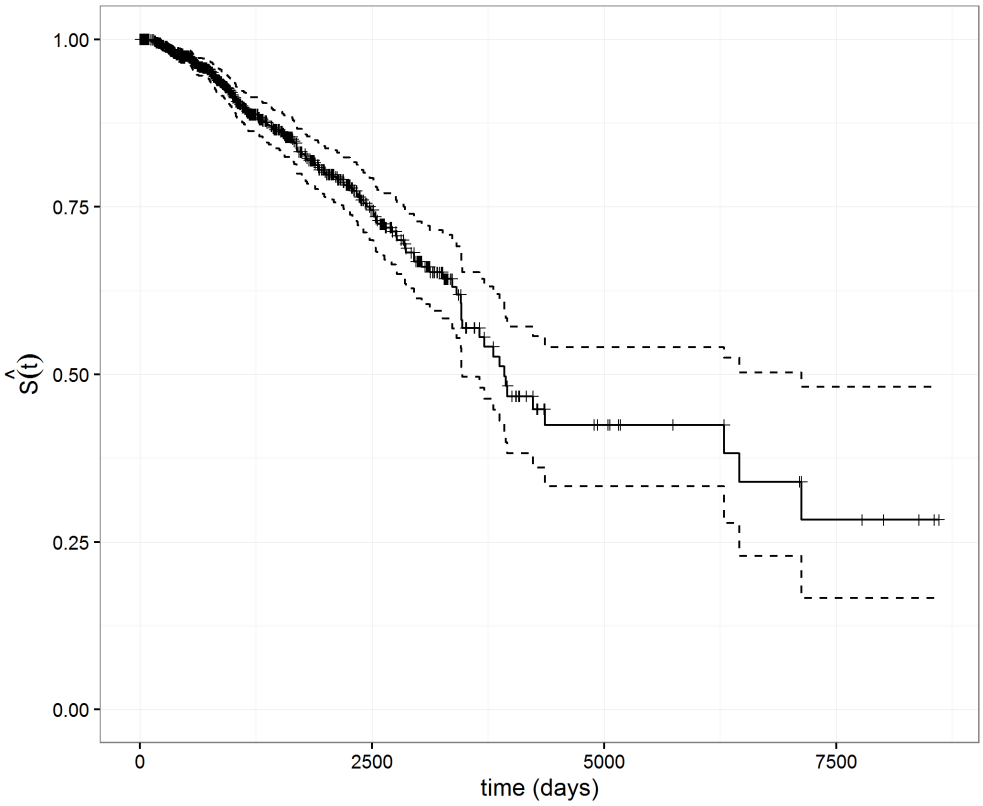

Figure 6. Kaplan-Meier estimate of the survivor function $S(\hat{t})$ of the original dataset, after removing male observations, denoted as the solid line. The 95% confidence interval is shown as the dotted band.

lymph nodes are integer valued, while the other covariates are categorical. Several issues with these covariates were identified and addressed. In all cases, the transformation or omission of a covariate does not drastically affect our analysis, as our goal does not involve interpretation of covariate effects, only a comparison of models which share clinical covariates.

"Pathologic_stage", which is an assessment of the severity of the cancer at the time of initial diagnosis, contains several levels where all patients survived. Fitting a Cox model on survival using this covariate is not possible due to numerical issues and a lack of convergence in the fitting procedure. We avoided this issue by creating a new categorical variable, *stage*, where the stages were collapsed down into five levels (one for each major stage of cancer, and a fifth for the cases where cancer stage was not determined or missing).

pathologic_stage	Censored	Deceased
stage i	77	13
stage ia	82	3
stage ib	6	0
stage ii	5	0
stage iia	319	34
stage iib	220	29
stage iii	0	2
stage iiia	129	26
stage iiib	18	8
stage iiic	53	9
stage iv	5	14
stage x	7	7
NA	6	2

stage	Censored	Deceased
1	165	16
2	544	63
3	200	45
4	5	14
unknown	13	9

Table 4. Counts of "pathologic_stage" and collapsed *stage* variables, tabulated against survival outcome. "NA" and "stage x" were collapsed into the "unknown" level.

"Race" yielded a similar problem – the only "american indian or alaska native" was censored, and a number of patients did not have a recorded race. We elected to omit "race" as a categorical covariate.

We ran a standard Cox model on the remaining clinical covariates;

"race"	Censored	Deceased
american indian or alaska native	1	0
asian	58	3
black or african american	151	29
white	630	108
NA	87	7

Table 5. Counts of the "race" variable tabulated against survival outcome.

"years_to_birth", "radiation_therapy", "number_of_lymph_nodes", and *stage*. The results suggest that "number_of_lymph_nodes" does not have a statistically significant effect on survival outcome when adjusting for the other covariates. This is likely due to the high correspondence between it and the *stage* variable. We omitted this covariate from further analysis as well. We used a likelihood ratio test to verify our decision to exclude "number_of_lymph_nodes". The fitted Cox model and likelihood ratio test are shown in Table 6.

	coef	exp(coef)	se(coef)	z	p
years_to_birth	0.03	1.03	0.01	3.41	0.00
radiation_therapyyes	-0.66	0.52	0.23	-2.84	0.00
number_of_lymph_nodes	0.03	1.03	0.02	1.44	0.15
stage2	0.75	2.11	0.41	1.80	0.07
stage3	1.67	5.30	0.45	3.67	0.00
stage4	2.44	11.53	0.56	4.38	0.00
stageunknown	0.89	2.43	0.81	1.10	0.27

	loglik	Chisq	Df	P(>|Chi|)
with lymph	-463.92			
without lymph	-464.89	1.93	1	0.1647

Table 6. Cox model estimates, along with Wald statistics and associated p-values for each coefficient where stage1 is the reference class, and Likelihood Ratio Test (LRT) of the model compared to a model with the lymph node covariate excluded. "number_of_lymph_nodes" is the only coefficient with a statistically insignificant p-value. The LRT suggests there is not a statistically significant difference between the two models.

Our final dataset consisted of survival outcome, "years_to_birth", "radiation_therapy", *stage*, and RSEM covariates for each patient. Patients with missing information were omitted. This left 967 observations, with 109 events and 858 cen-

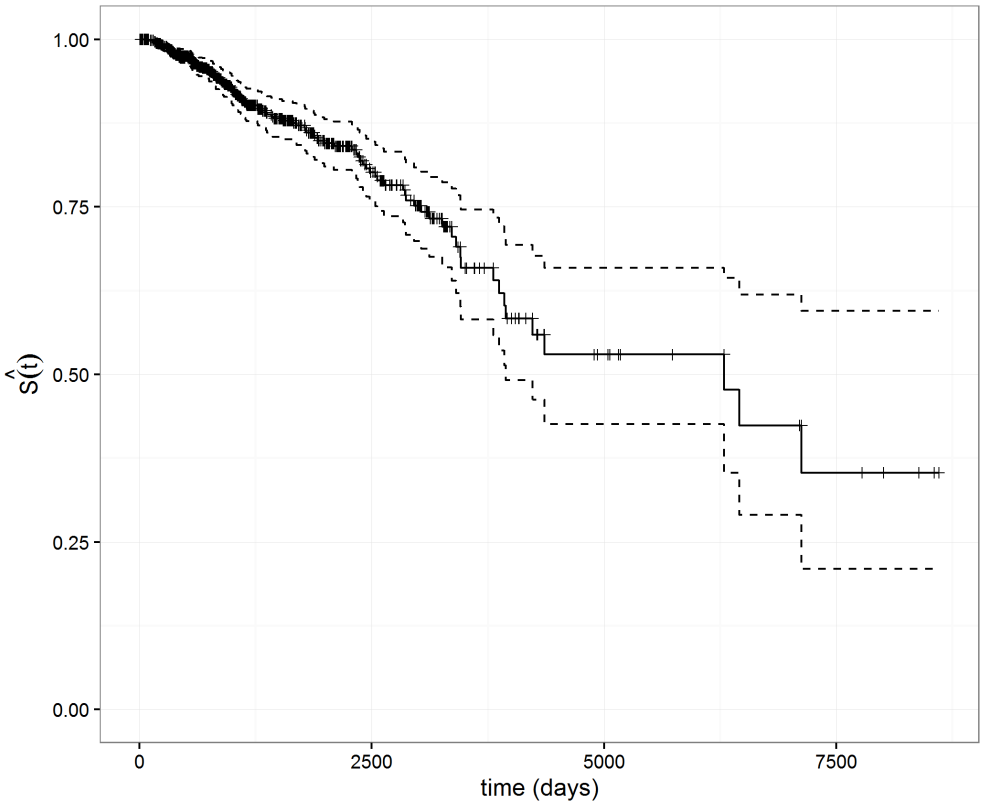

Figure 7. Kaplan-Meier estimate of the survivor function $S(\hat{t})$ for the merged data denoted as the solid line. The 95% confidence interval is shown as the dotted band. The removal of additional observations (mostly censored) has raised the curve and increased the median observed survival time.

sored cases. The median survival time increased to 6456 days. The Kaplan-Meier survival curve is shown in Figure 7.

2.2 Filtering

Although the methods we use during the analysis fit over-parameterized models, the resulting covariate effect estimates are still limited by the number of observations. In particular, the lasso estimator will return at most 967 coefficient estimates. This is just over 1% of the isoform covariates! While the ridge estimator will return coefficient estimates for all covariates, it is unlikely that these will represent the true effects due to the lack of power. In order to increase the chances of fitting models with meaningful predictive power, we reduced the number of isoform covariates by con-

ducting per-variable selection procedure while correcting for multiple testing using a local false discovery rate (fdr) threshold.

2.2.1 Per-Variable Models

We fit a standard Cox model for each individual isoform which has a known correspondence to a gene (63,213 isoforms), where each model contained only the single isoform covariate. As is common in genomic analysis of CPMs, the base-2 logarithmic transformation was taken on each covariate.

For each model, we collected the probability of estimating a covariate effect at least as large as the model's estimated effect under a null hypothesis where the true covariate effect is 0 (the p-value for the covariate). This was done using a Likelihood Ratio Test (LRT): the scaled difference between log likelihoods of models with and without the effect comes from a χ^2 distribution with one degree of freedom if the models are equivalent, as the sample size goes to infinity (Wilks, 1938).

These are plotted in Figure 8 against the theoretical distribution of p-values under the null hypothesis, i.e. a Uniform 0,1 distribution. The deviation between the observed p-value distribution and the theoretical distribution suggests there are isoform covariates which do have a nonzero effect on patient survival time despite the large number of tests (and subsequent inflation of the overall Type 1 error rate). Some of the counts from Figure 8 are broken out in Table 7.

# models	63213
# non-NA	63213
# <= .05	5417
# <= .001	215

Table 7. Counts of per-isoform Cox models and models with extremely low p-values. Isoform filtering is based on a later correction to meet a false discovery rate threshold due to multiple testing.

We aggregated the isoform counts into gene-level counts, yielding 19,330 gene

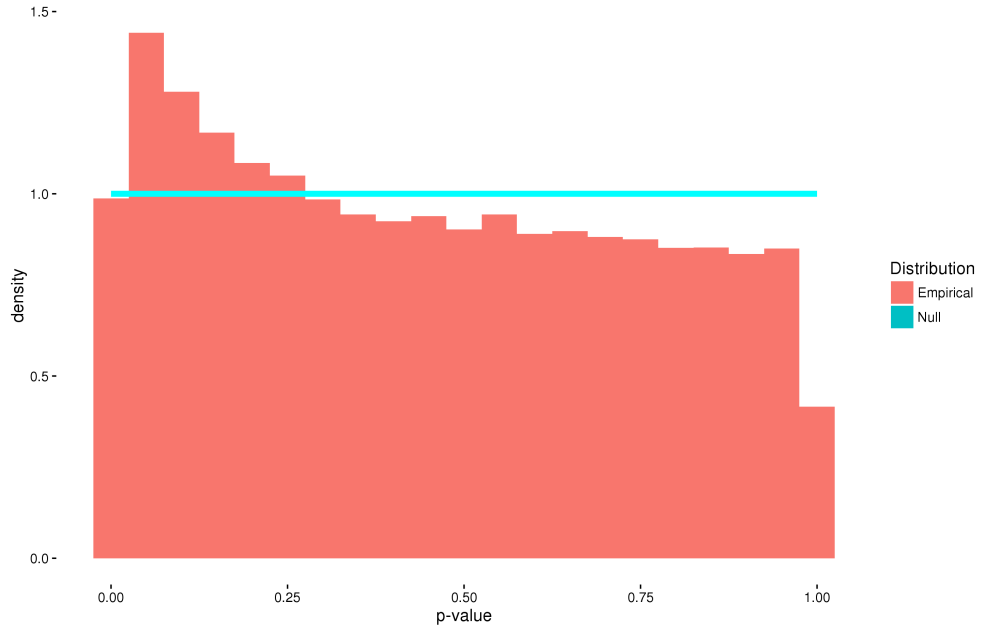

Figure 8. Relative histogram of binned p-values of isoform covariates in univariate per-isoform Cox models, compared to the theoretical distribution under the null assumption of no effect on survival time. The difference in distributions suggests the null assumption does not hold.

level covariates. We then repeated the same univariate modeling procedure, yielding Figure 9. Strangely enough, several of the isoform and gene aggregated covariates were duplicated among patients. The full list of genes which had at least one duplicated isoform among all patients is given in Table 8.

ACSBG2|81616	MCCC1|56922	RBMY1J|378951
AK2|204	NR1I3|9970	RBMY2FP|159162
HNRNPC|3183	POLDIP3|84271	SEPT7|989
KDM5D|8284	RBMY1A1|5940	TMEM161A|54929
		VDAC3|7419

Table 8. Gene names where at least one constituent isoform had expression estimates duplicated across all patients.

2.2.2 Multiple Testing Correction

As alluded to in the previous section, conducting 63,213 or 19,330 tests (in the case of isoform or gene-level testing, respectively) constitutes a multiple testing is-

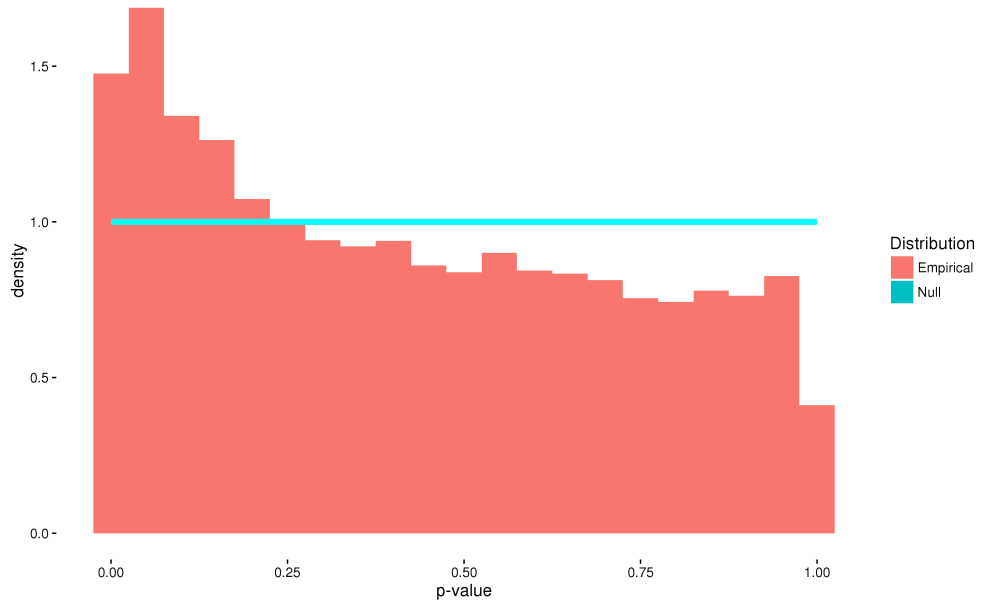

Figure 9. Relative histogram of binned p-values of gene-level covariates in univariate per-aggregate Cox models, compared to the theoretical distribution under the null assumption of no effect on survival time. The resulting plot is similar to Figure 8.

sue. Under the assumption of isoform independence and the null hypothesis, we'd expect to see 3160 isoforms with a p-value less than 0.05. That is, even if all isoform expression levels have no effect on patient survival time (assuming all isoform expression levels are independent of one another) we'd likely find a large number spurious correlations between expression and survival time relative to the number of patients in our dataset.

One approach to identifying and separating the "interesting" isoforms or genes from those with spurious correlations is to control the local false discovery rate (Efron and Tibshirani, 2002). We assume there are two underlying classes of covariates ("interesting" and "unimportant") and that each isoform or gene covariate belongs to one of the two classes. "Unimportant" covariates are assumed to have no "true" effect on patient survival time, while "interesting" covariates have a nonzero effect. We assume each covariate has a η_0 probability of being "unimportant" and a

$1-\eta_0$ probability of being "interesting". If the "unimportant" and "interesting" covariates are identically distributed within their classes, the p-values, denoted y, from the per-covariate models follow a mixture density

$$f(y) = \eta_0 f_0(y) + (1-\eta_0) f_1(y)$$

where f_0 is the density function of the "unimportant" p-values and f_1 is the density function of the "interesting" p-values. Since f_0 is the density function of the null hypothesis, we can define the probability of a covariate being "unimportant" given its p-value equals y as

$$\text{fdr}(y) = \frac{\eta_0 f_0(y)}{f(y)}$$

which is the definition of the local false discovery rate, or fdr. Here, f_0 is a Uniform 0,1 density. We use the R package "fdrtool" (Strimmer, 2008) to estimate both $\hat{\eta_0}$ and $\hat{f(y)}$ and construct the set of fdr probabilities for each isoform and gene-level covariate.

$\hat{\eta_0}$ was estimated at 0.8408 for the isoform mixture and 0.7772 for the gene-level mixture. This makes sense; there is likely a higher proportion of "noise" covariates at the isoform level than when they are collapsed down into aggregates representing an overall gene level.

We plotted an unscaled version of the empirical cumulative distribution function (CDF) for the estimated fdr values in Figures 10 and 11.

We used a cutoff threshold of 0.20; that is, accepting covariates with an estimated probability of being "unimportant" given their p-value – a false discovery – less than 0.20. This yields 81 isoforms from the per-isoform model p-values and 298 genes from the per-gene model p-values.

The 81 isoforms with estimated fdr values less than or equal to 0.20 correspond to 76 unique genes. To make a fair comparison between isoform-level and gene-level models in the analysis, we consider the set of 76 unique gene covariates against their

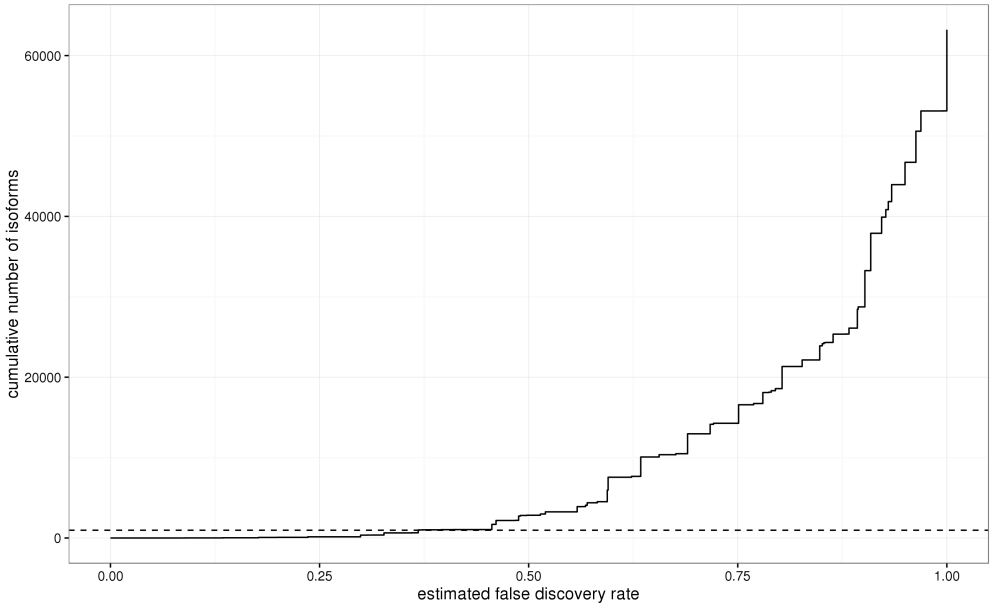

Figure 10. Empirical CDF of per-isoform estimated fdr values. The dashed line represents the 967 observations and meets the CDF at a cutoff value of 0.367. This is a rough indication of when the number of covariate effects to be estimated will be greater than the number of observations.

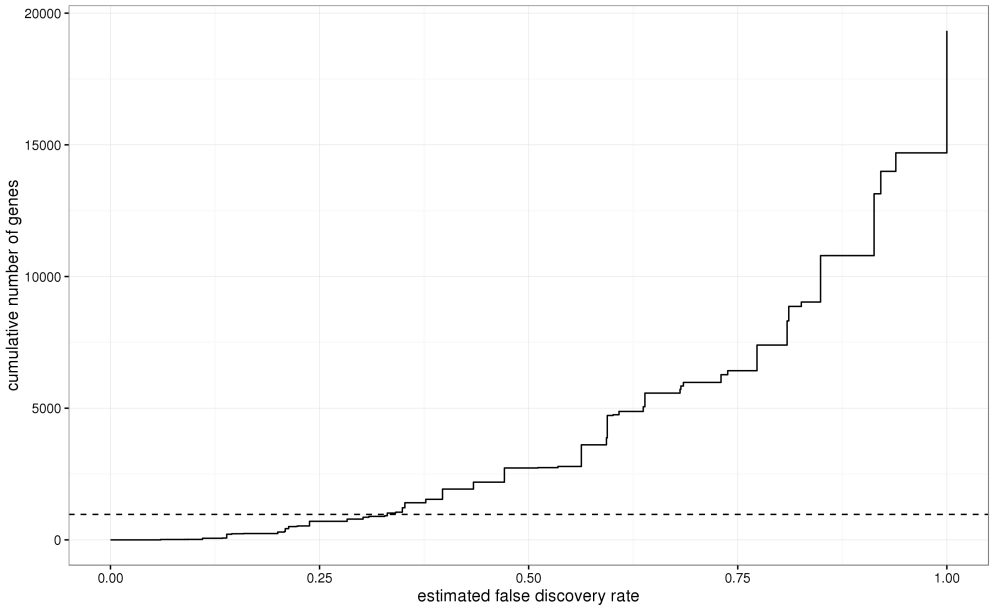

Figure 11. Empirical CDF of per-gene estimated fdr values. The dashed line representing the 967 observations meets the CDF at a cutoff value of 0.330.

constituent 332 isoform variants (including the 81 "interesting" isoforms). The 298 "interesting" genes correspond to 914 isoform variants.

2.3 Analysis

We fit a battery of penalized models to the merged and filtered dataset. Two overall sets of models were fit to the data. Both sets of models were fit with the high-level clinical covariates. The first set of models included the genetic covariates determined by the isoforms with significant univariate association with overall survival after controlling for a false discovery rate of 0.20. The second set included the genetic covariates based on gene-level aggregates with significant associations with overall survival after controlling for the same false discovery rate. In all cases, the log base2 transformation of the CPM was used.

2.3.1 Cross-Validated Loss

We performed a grid search to assess the impact of α and λ on the fit of the data. α is the model parameter which determines the proportion of the L_1 and L_2 norms in the penalty term of the elastic net, where $\alpha = 0$ corresponds to the LASSO and $\alpha = 1$ corresponds to ridge regression. λ is the model parameter which determines the magnitude of the penalty term, where $\lambda = 0$ corresponds to an unpenalized model.

Quality of fit was determined using the cross-validated loss procedure of (Simon et al., 2011) as implemented in the package "glmnet" (Friedman et al., 2013). The goodness-of-fit for a model is given by splitting the dataset into a series of "folds", and estimating the model parameters for each partition of the dataset which excludes a "fold". The difference between the partial-likelihood of the estimated parameters on the full dataset and the partial-likelihood of the estimated parameters on the partition is taken, and the sum of the differences is the cross-validated loss. The values of λ which minimize the cross-validated loss

are considered to "best" fit the data.

However, if the number of folds is less than the number of observations, the partitioning of the data is not necessarily unique and the cross validation procedure should be repeated to assess the impact of this variability. We opted to eliminate this variability by finding the cross-validated loss using leave-one-out cross validation (LOOC) where partition contained all but one observations. This guarantees that each partition is unique and that the estimated model parameters found will not vary.

2.3.2 Concordance

We computed the c-statistics using the procedure implemented in the package "survival" (Therneau, 2016) for the subset of the models in our cross validated grid search which had the "best" values of λ for a given α parameter.

We also compared isoform and gene-aggregate models using the variant of the c-statistic implemented in the package "survC1" (Uno, 2013). We provided the risk scores for the cross validated models specified above as univariate predictors and computed the concordance. We then tested the difference in concordance for statistical significance between the model corresponding to the gene-aggregate covariates and the model corresponding to their constituent isoform covariates, given a fixed value of α and each model's "best" value of λ according to the cross-validation procedure. This procedure was run for both sets of covariates, those determined by isoform-level filtering as well as those determined by gene-level filtering.

List of References

Broad Institute TCGA Genome Data Analysis Center (2016). Analysis-ready standardized tcga data from broad gdac firehose 2016_01_28 run.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):1.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86.

Friedman, J., Hastie, T., and Tibshirani, R. (2013). glmnet: Lasso and elastic-net regularized generalized linear models. version1.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2011). Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, page kxr031.

Pimentel, H. (2014). In rna-seq, 2 != 2: Between-sample normalization. [Online]. Available: https://haroldpimentel.wordpress.com/2014/12/08/in-rna-seq-2-2-between-sample-normalization/. Accessed: 08/20/2016.

R Development Core Team. R: A language and environment for statistical computing. version 3.2.5.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1.

Strimmer, K. (2008). fdrtool: a versatile r package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461–1462.

Therneau, T. (2016). A package for survival analysis in s.

Uno, H. (2013). Package 'survc1'.

VanRossum, G. and Drake, F. L. (2010). *The Python Language Reference*. Python software foundation Amsterdam, Netherlands.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

CHAPTER 3

Results and Conclusions

3.1 Cross-validated Model Fitting

We present the results of the cross-validated loss procedure and discuss the fits and number of nonzero coefficient estimates for isoform and gene-aggregate models.

Cross-validated loss per observation, $\hat{CV}_i$, as described previously (2.3.1), is defined mathematically as a function of λ and α:

$$\hat{CV}_i(\lambda,\alpha) = \ell(\hat{\beta_{-i}}(\lambda,\alpha)) - \ell_{-i}(\hat{\beta_{-i}}(\lambda,\alpha))$$

where ℓ is the partial-likelihood function of the complete dataset, ℓ_{-i} is the partial-likelihood function of the dataset excluding the ith observation, and $\hat{\beta_{-i}}$ are the coefficient estimates for the dataset excluding the ith observation. We used leave-one-out cross-validated (LOOC) loss, the mean over $\hat{CV}_i$ for all i observations, to compare isoform and gene-aggregate models. We sometimes refer to LOOC loss as the overall cross-validated loss.

The values of overall cross-validated loss are show in Tables 9 and 10 for each value of α in our grid search and the corresponding value of λ which minimized the loss.

3.1.1 Isoform-based Models

Table 9 refers to models fit with the 332 isoform covariates or 76 gene-aggregate covariates, along with the selected clinical variables. These versions of the covariates correspond to the set of isoforms which were found to have high levels of agreement with survival outcome during the univariate filtering procedure. The form of the covariates (whether broken out by isoform or lumped together as a gene-aggregate) is denoted by the "type" column.

type	alph	lambda	cvm	cvsd	nzero	gzero	percnzero
iso	0.00	0.14	10.00	0.08	338	76	1.00
gene	0.00	0.11	10.58	0.07	82	76	1.00
iso	0.10	0.06	10.10	0.08	171	69	0.91
gene	0.10	0.05	10.57	0.07	56	51	0.67
iso	0.20	0.04	10.15	0.08	131	62	0.82
gene	0.20	0.04	10.58	0.07	49	44	0.58
iso	0.30	0.03	10.18	0.09	113	58	0.76
gene	0.30	0.03	10.59	0.07	44	39	0.51
iso	0.40	0.02	10.21	0.09	104	57	0.75
gene	0.40	0.02	10.59	0.07	40	35	0.46
iso	0.50	0.02	10.23	0.09	94	54	0.71
gene	0.50	0.02	10.59	0.07	38	33	0.43
iso	0.60	0.02	10.25	0.09	92	53	0.70
gene	0.60	0.02	10.60	0.07	37	32	0.42
iso	0.70	0.02	10.27	0.08	78	48	0.63
gene	0.70	0.02	10.60	0.07	35	30	0.39
iso	0.80	0.02	10.29	0.08	69	44	0.58
gene	0.80	0.01	10.60	0.07	35	30	0.39
iso	0.90	0.02	10.29	0.08	69	44	0.58
gene	0.90	0.01	10.59	0.07	36	31	0.41
iso	1.00	0.01	10.30	0.08	69	44	0.58
gene	1.00	0.01	10.59	0.07	33	28	0.37

Table 9. Summary of cross-validated loss in models fit to the covariates determined by univariate filtering at the isoform level. For each given level of α, the model with the lowest value of mean cross-validated loss ("cvm") over all possible values of λ is shown. Models run on the 332 constituent isoforms are labeled as "iso" type, and those run on the 76 gene-level aggregates are labeled as "gene" type. The standard deviation ("cvsd") and number of non-zero model coefficient estimates ("nzero") are displayed. For comparison purposes, we've added the number and proportion of the 76 genes represented by at least one non-zero isoform or aggregate coefficient estimate in the "gzero" and "percnzero" columns.

The cross validated loss ("cvm" column) for all of the 11 penalized models is lower for models fit on the "iso" type of covariates and the lowest loss occurs when using the ridge estimator ($\alpha = 0$). The ridge estimator does not perform variable selection and thus produces non-zero coefficient estimates for all 332 isoform covariates or all 76 gene-aggregate covariates (as well as the 6 clinical variables) as shown in the "nzero" column. As more weight is placed on the L_1 norm in the penalty term, the number of variables in the model drops to 69 and 33 respectively when using the LASSO estimator ($\alpha = 1$). We've added a column representing the number of the 76 genes which are represented by at least one non-zero coefficient estimate ("gzero") to facilitate comparison between the variable selections in both types of models. The models fit on the "iso" type of covariates always represent variants of the same or more unique genes than those fit on the "gene" type of covariates and even the LASSO estimator produces a model with over half of the genes represented. The proportion of genes with non-zero coefficient estimates is shown in the "percnzero" column.

The grid search over λ and α on these covariates is denoted in Figures 12 and 13. The models which minimize the loss tend to have a number of null or zero coefficient estimates (with the exception of the ridge estimator model), even for low values of α. As α decreases, the value of λ for the model with the lowest loss increases.

3.1.2 Gene-based Models

Table 10 refers to models fit with the 914 isoform covariates or 298 gene-aggregate covariates selected by univariate filtering on the gene-aggregate covariates, along with the selected clinical variables.

The relationship of cross validated loss ("cvm" column) for all of the 11 penalized models is reversed from the previous section; for any given value of α, the loss is lower for models fit on the "gene" type of covariates. As before, the lowest loss occurs when using the ridge estimator ($\alpha = 0$). The number of covariates removed as α increases

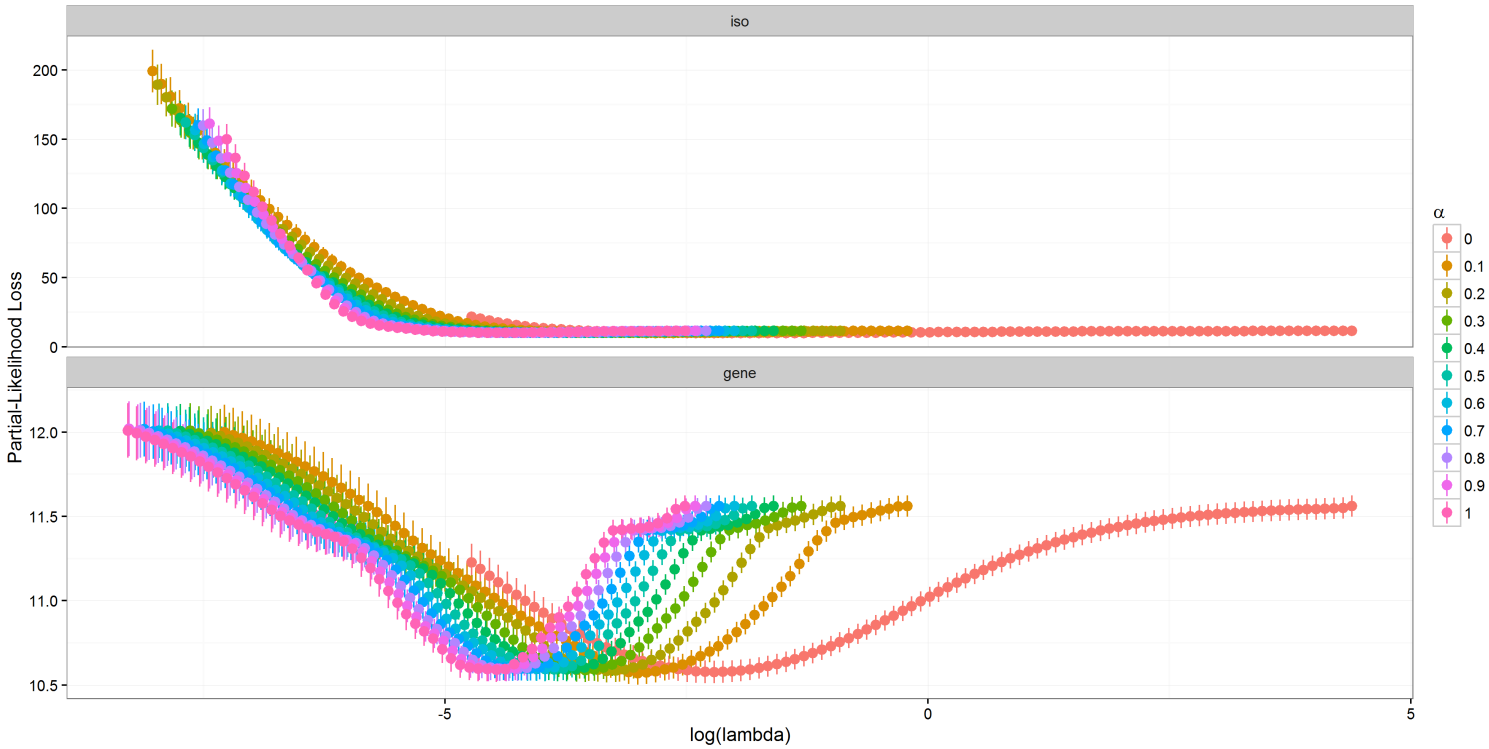

Figure 12. Partial Likelihood Loss as a function of $\log(\lambda)$ on the covariates determined by univariate filtering at the isoform level. The top plot refers to models fit with the "iso" covariates (332 isoforms) while the bottom plot refers to those fit with the "gene" covariates (76 gene-level aggregates). α values of 0 and 1 are the ridge and LASSO estimator, respectively.

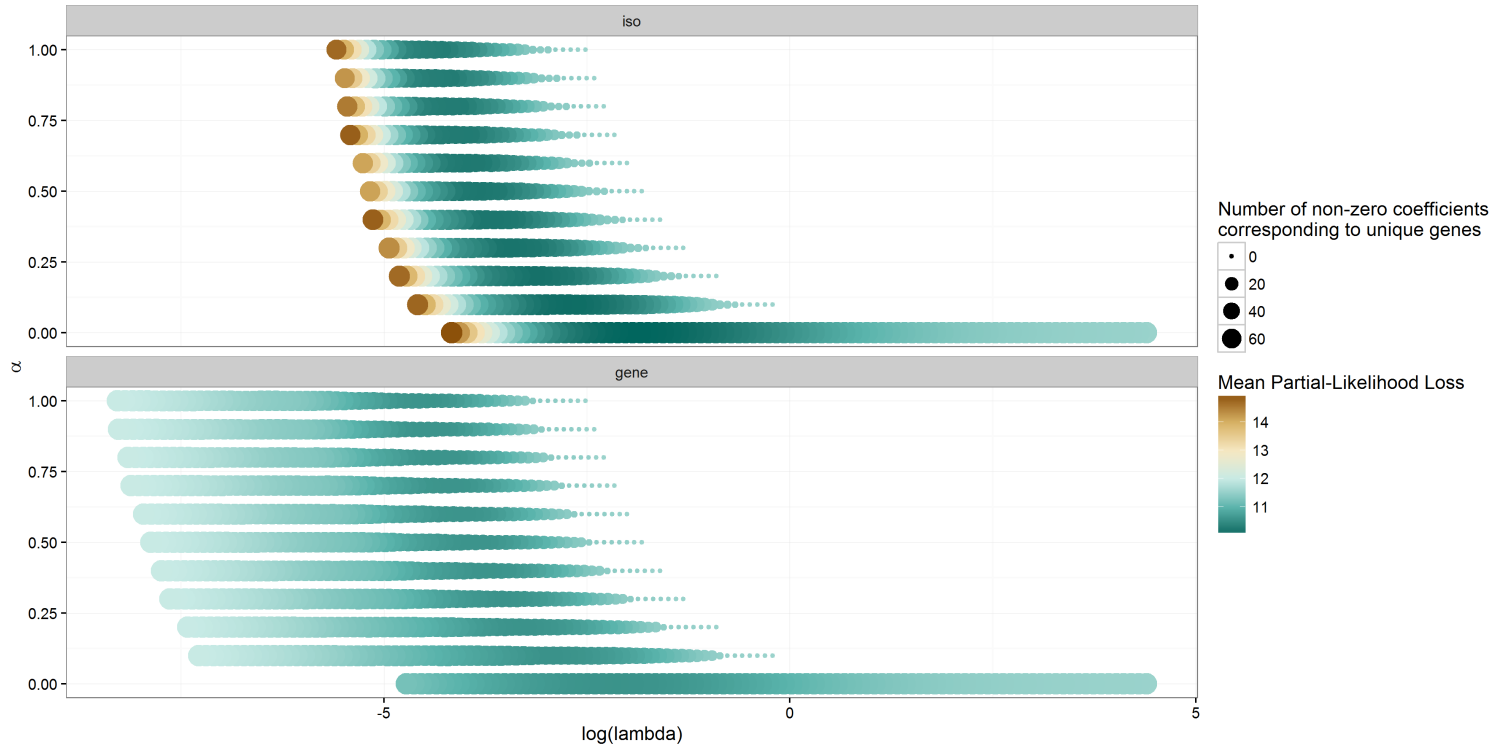

Figure 13. Models with partial likelihood losses of less than 15, as a function of $\log(\lambda)$ and α on the covariates determined by univariate filtering at the isoform level. The top plot refers to models fit with the "iso" covariates (332 isoforms) while the bottom plot refers to those fit with the "gene" covariates (76 gene-level aggregates). α values of 0 and 1 are the ridge and LASSO estimator, respectively. The size of the points represents the number of genes with non-zero coefficients out of the 76 total. Nearly null models (with no genetic covariates) can be seen as λ increases and the number of non-zero coefficients plummets.

type	alph	lambda	cvm	cvsd	nzero	gzero	percnzero
iso	0.00	0.33	10.45	0.08	920	298	1.00
gene	0.00	0.14	10.37	0.08	304	298	1.00
iso	0.10	0.09	10.55	0.08	261	170	0.57
gene	0.10	0.05	10.45	0.08	141	135	0.45
iso	0.20	0.05	10.59	0.08	190	133	0.45
gene	0.20	0.04	10.46	0.07	92	87	0.29
iso	0.30	0.05	10.62	0.07	125	98	0.33
gene	0.30	0.03	10.48	0.07	80	75	0.25
iso	0.40	0.03	10.62	0.07	115	90	0.30
gene	0.40	0.03	10.52	0.07	75	70	0.23
iso	0.50	0.03	10.64	0.08	112	88	0.30
gene	0.50	0.02	10.55	0.07	69	64	0.21
iso	0.60	0.02	10.66	0.08	119	92	0.31
gene	0.60	0.02	10.56	0.07	68	63	0.21
iso	0.70	0.02	10.67	0.08	115	88	0.30
gene	0.70	0.02	10.57	0.07	64	59	0.20
iso	0.80	0.02	10.68	0.07	79	64	0.21
gene	0.80	0.01	10.58	0.07	64	59	0.20
iso	0.90	0.02	10.67	0.07	77	63	0.21
gene	0.90	0.01	10.59	0.07	63	58	0.19
iso	1.00	0.02	10.67	0.07	73	61	0.20
gene	1.00	0.01	10.59	0.07	62	57	0.19

Table 10. Summary of cross-validated loss in models fit to the covariates determined by univariate filtering at the gene-aggregate level. For each given level of α, the model with the lowest value of mean cross-validated loss ("cvm") over all possible values of λ is shown. Models run on the 914 constituent isoforms are labeled as "iso" type, and those run on the 298 gene-level aggregates are labeled as "gene" type. The standard deviation ("cvsd") and number of non-zero model coefficient estimates ("nzero") are displayed. For comparison purposes, we've added the number and proportion of the 298 genes represented by at least one non-zero isoform or aggregate coefficient estimate in the "gzero" and "percnzero" columns.

and more weight is placed on the L_1 norm is also much greater than in the previous section, though this is likely an artifact of the larger number of covariates overall.

The column representing the number of the 298 genes which are represented by at least one non-zero coefficient estimate ("gzero") shows that, as in the previous section, models fit on the "iso" type of covariates always represent the same or more unique genes than those fit on the "gene" type of covariates.

The grid search over λ and α on these covariates is denoted in Figures 14 and 15.

3.1.3 Comparison

The relationship between the models fit on the isoform-filtering covariates and the gene-filtering covariates is shown in Figure 16. The reversed relationship between "type" and "cvm" is visible. The cross-validated loss is lowest for models fit on isoform covariates determined by isoform-level univariate filtering, but is highest for models fit on isoform covariates determined by gene-level univariate filtering.

3.2 Concordance

We present high level results of concordance measures for the models presented in the previous section. Concordance was computed on the observations of the models fit above. Rationale and shortcomings of this approach are treated in the discussion.

3.2.1 C-Statistics

The c-statistic, as defined previously in 1.4.2, of the linear predictor, $\hat{\beta}X_i$, for all the models in Tables 9 and 10, is given in Tables 11 and 12.

The relationship between the "iso" and "gene" models remains consistent between both sets of covariates. For both sets of covariates, "iso" type models have a higher point estimate of concordance than "gene" type models. The most concordant models are often those fit using the ridge estimator. Both sets of covariates show

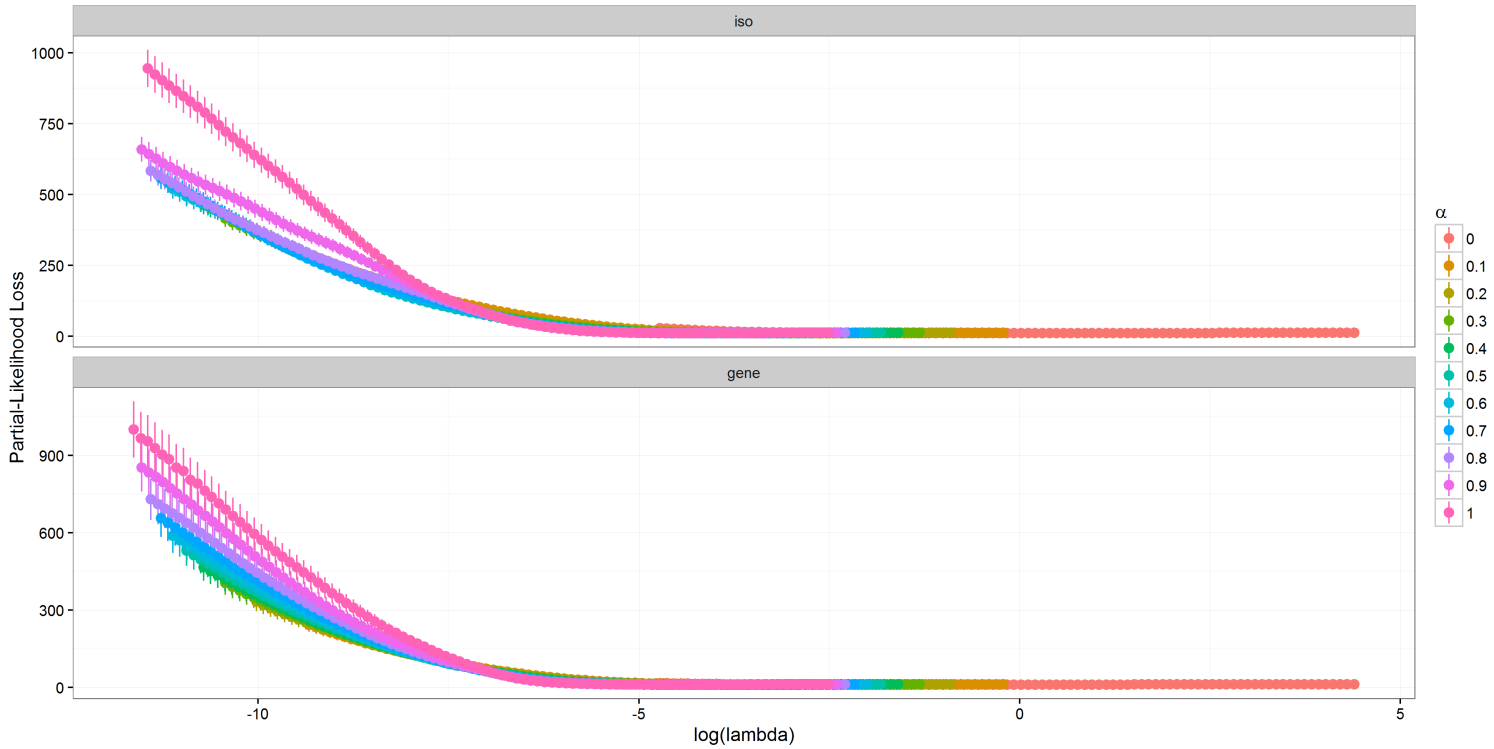

Figure 14. Partial Likelihood Loss as a function of $\log(\lambda)$ on the covariates determined by univariate filtering at the gene-aggregate level. The top plot refers to models fit with the "iso" covariates (914 isoforms) while the bottom plot refers to those fit with the "gene" covariates (298 gene-level aggregates). α values of 0 and 1 are the ridge and LASSO estimator, respectively.

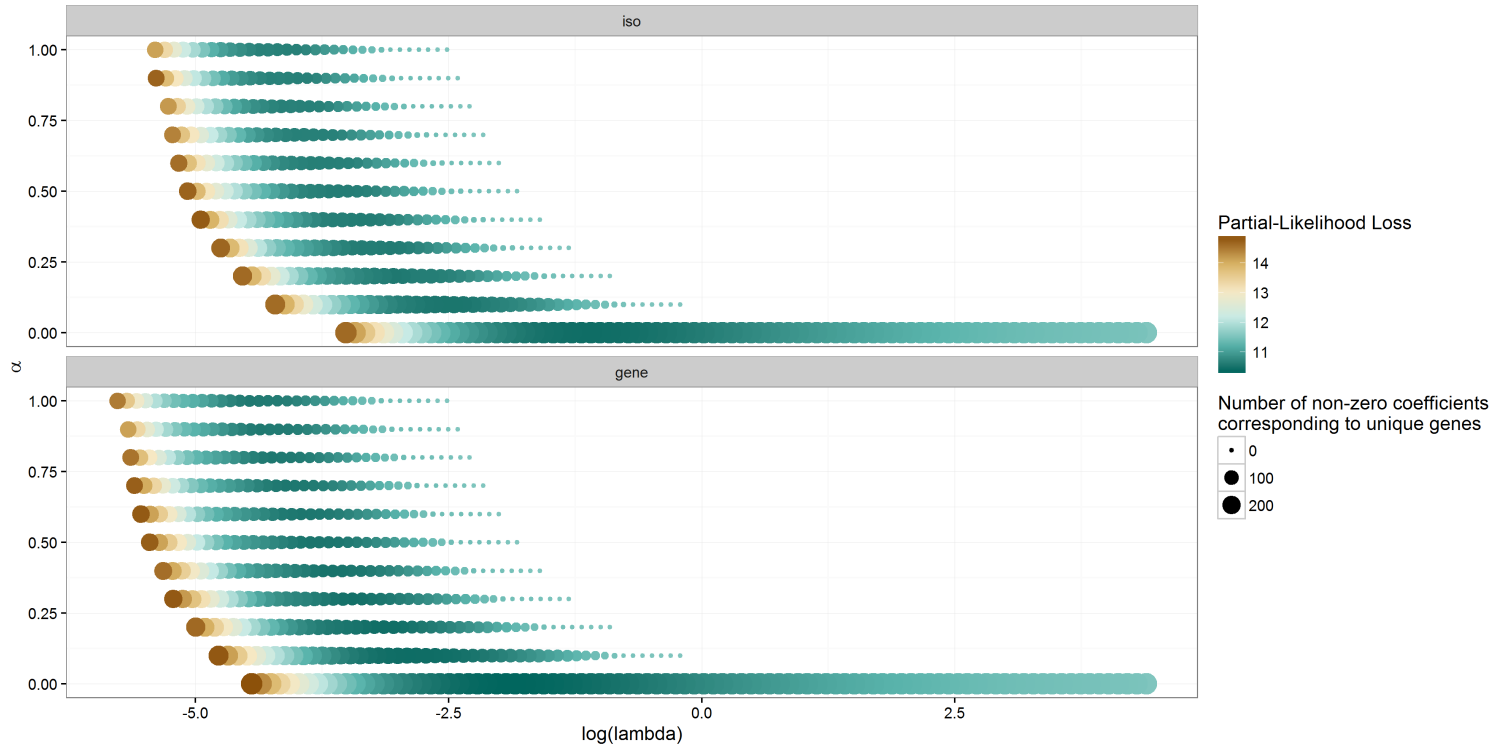

Figure 15. Models with partial likelihood losses of less than 15, as a function of $\log(\lambda)$ and α on the covariates determined by univariate filtering at the gene-aggregate level. The top plot refers to models fit with the "iso" covariates (914 isoforms) while the bottom plot refers to those fit with the "gene" covariates (298 gene-level aggregates). α values of 0 and 1 are the ridge and LASSO estimator, respectively. The size of the points represents the number of genes with non-zero coefficients out of the 298 total. Nearly null models (with no genetic covariates) can be seen as λ increases and the number of non-zero coefficients plummets.

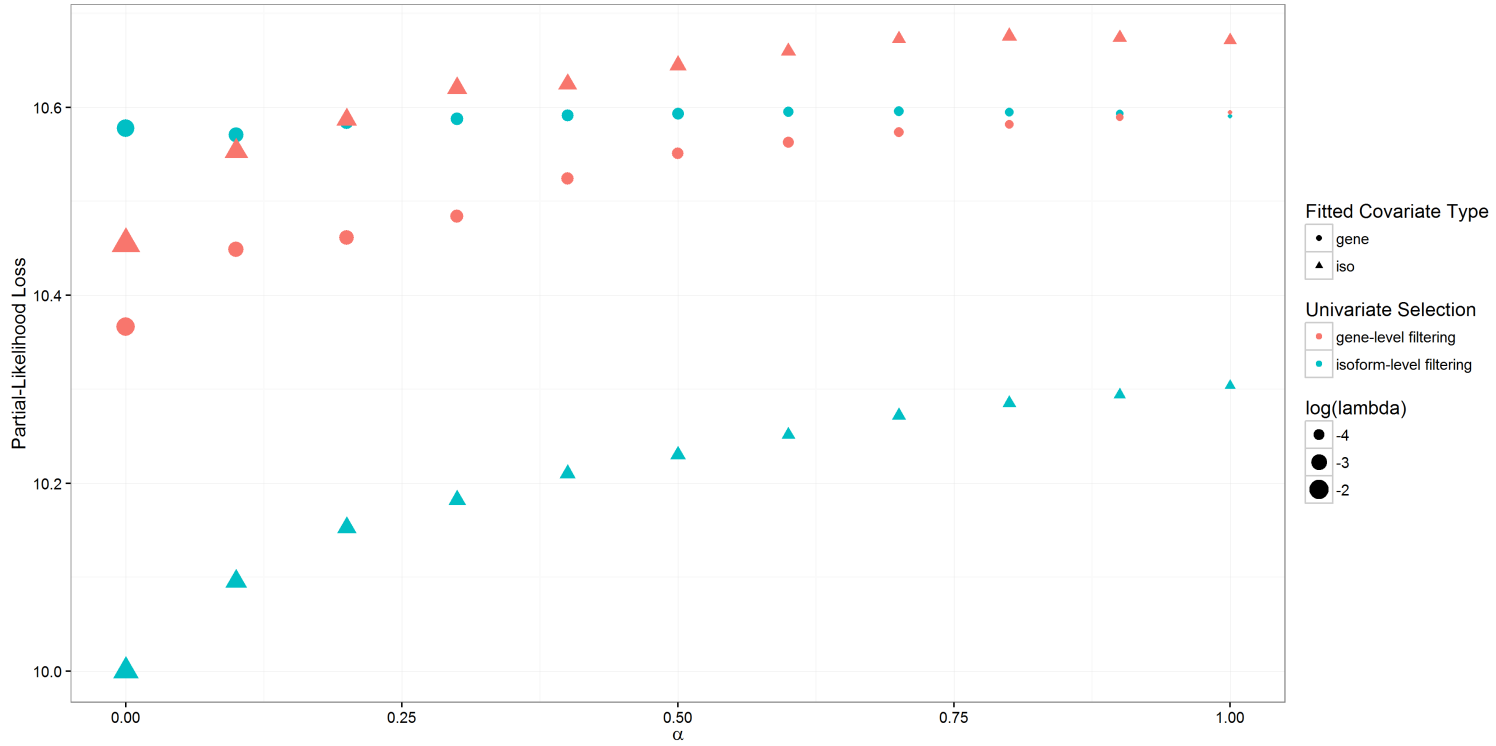

Figure 16. Models from Tables 9 and 10 plotted on the same set of axes, α versus "cvm". The "type" column is denoted by shape, and Table 9's models appear in blue, while Table 10's models are plotted in red. Each model's value of $\log(\lambda)$ is denoted by the size of its mark. The ridge estimator models appear on the far left side of the plot with the largest λ values, while the LASSO models are on the rightmost side.

similar high concordance along the range of α values when used in "iso" type models.

type	alph	lambda	c-statistic
iso	0.00	0.14	0.95
gene	0.00	0.11	0.87
iso	0.10	0.06	0.94
gene	0.10	0.05	0.87
iso	0.20	0.04	0.94
gene	0.20	0.04	0.86
iso	0.30	0.03	0.94
gene	0.30	0.03	0.86
iso	0.40	0.02	0.94
gene	0.40	0.02	0.86
iso	0.50	0.02	0.93
gene	0.50	0.02	0.86
iso	0.60	0.02	0.94
gene	0.60	0.02	0.86
iso	0.70	0.02	0.93
gene	0.70	0.02	0.86
iso	0.80	0.02	0.92
gene	0.80	0.01	0.86
iso	0.90	0.02	0.92
gene	0.90	0.01	0.86
iso	1.00	0.01	0.92
gene	1.00	0.01	0.86

Table 11. C-statistics for each model from Table 9. These models were fit to the set of covariates determined by univariate filtering at the isoform-level. "iso" models were fit to the set of 332 constituent isoforms and "gene" models were fit to 76 gene-level aggregates.

3.2.2 Uno's C

We used Uno's C estimator to re-estimate concordance and obtain an estimate of the standard error. Uno's C (along with standard errors) are shown graphically against the c-statistics in Figure 17. Since the sampling distribution of Uno's C is asymptotically normal, we constructed confidence intervals for the difference in concordance between "iso" and "gene" type models and present the results in Tables 13 and 14. The differences are also shown graphically in Figure 18.

type	alph	lambda	concordance
iso	0.00	0.33	0.96
gene	0.00	0.14	0.92
iso	0.10	0.09	0.95
gene	0.10	0.05	0.92
iso	0.20	0.05	0.95
gene	0.20	0.04	0.91
iso	0.30	0.05	0.93
gene	0.30	0.03	0.91
iso	0.40	0.03	0.93
gene	0.40	0.03	0.91
iso	0.50	0.03	0.93
gene	0.50	0.02	0.91
iso	0.60	0.02	0.94
gene	0.60	0.02	0.91
iso	0.70	0.02	0.94
gene	0.70	0.02	0.91
iso	0.80	0.02	0.92
gene	0.80	0.01	0.91
iso	0.90	0.02	0.92
gene	0.90	0.01	0.91
iso	1.00	0.02	0.92
gene	1.00	0.01	0.91

Table 12. C-statistics for each model from Table 10. These models were fit to the set of covariates determined by univariate filtering at the isoform-level. “iso” models were fit to the set of 914 constituent isoforms and “gene” models were fit to 298 gene-level aggregates.

alph	concordance	stderr	lower95	upper95
0.00	-0.11	0.02	-0.15	-0.07
0.10	-0.11	0.02	-0.15	-0.07
0.20	-0.11	0.02	-0.15	-0.07
0.30	-0.11	0.02	-0.15	-0.07
0.40	-0.11	0.02	-0.15	-0.07
0.50	-0.11	0.02	-0.14	-0.07
0.60	-0.11	0.02	-0.15	-0.07
0.70	-0.10	0.02	-0.14	-0.06
0.80	-0.09	0.02	-0.13	-0.05
0.90	-0.09	0.02	-0.13	-0.05
1.00	-0.09	0.02	-0.13	-0.06

Table 13. Difference in Uno's C between each "gene" and "iso" pair from Table 11. The difference is given in the "concordance" column. A negative difference indicates Uno's C is higher for the "iso" level model. The standard error of the difference and a 95% confidence interval based on the asymptotic distribution of the difference are reported as well.

alph	concordance	stderr	lower95	upper95
0.00	-0.06	0.01	-0.09	-0.03
0.10	-0.04	0.01	-0.07	-0.02
0.20	-0.05	0.01	-0.08	-0.03
0.30	-0.03	0.01	-0.05	-0.01
0.40	-0.03	0.01	-0.05	-0.01
0.50	-0.04	0.01	-0.06	-0.02
0.60	-0.05	0.01	-0.07	-0.02
0.70	-0.05	0.01	-0.07	-0.02
0.80	-0.02	0.01	-0.04	-0.00
0.90	-0.02	0.01	-0.04	-0.00
1.00	-0.02	0.01	-0.04	0.00

Table 14. Difference in Uno's C between each "gene" and "iso" pair from Table 12. The difference is given in the "concordance" column. A negative difference indicates Uno's C is higher for the "iso" level model. The standard error of the difference and a 95% confidence interval based on the asymptotic distribution of the difference are reported as well.

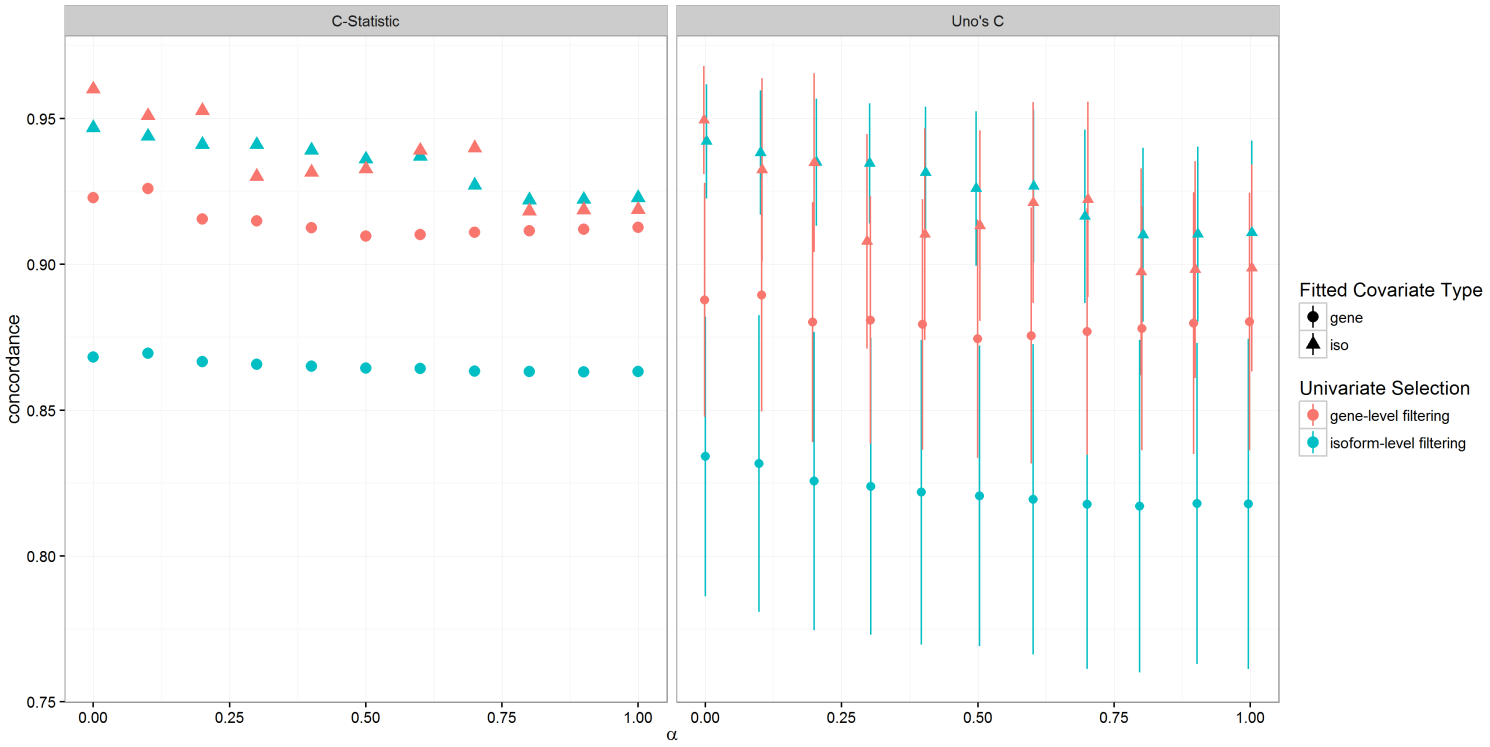

Figure 17. C-Statistics and Uno's C for models from Tables 9 and 10 plotted on the same set of axes, α versus "concordance". The "type" column is denoted by shape, and Table 9's models appear in blue, while Table 10's models are plotted in red. The estimated 95% confidence interval for each value of Uno's C, based on its asymptotic normality, is given by the line surrounding each point. Although the values differ between the estimators, the point estimates of the c-statistics lie within the estimated confidence intervals of Uno's C. The points are slightly jittered to show overlap in confidence intervals.

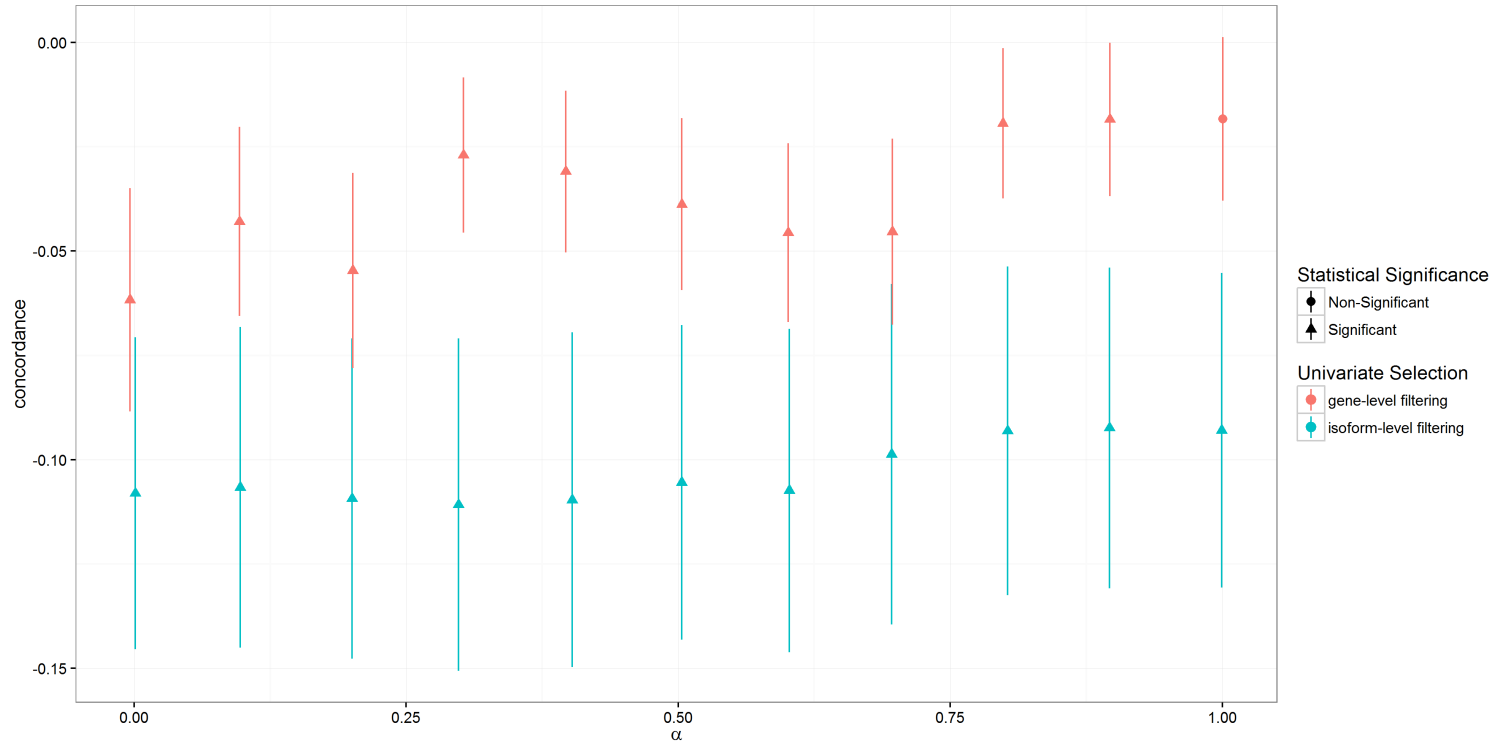

Figure 18. Difference in Uno's C for models from Tables 13 and 14 plotted on the same set of axes, α versus difference in "concordance". Table 13's models appear in blue, while Table 14's models are plotted in red. The estimated 95% confidence interval for each difference is given by the line surrounding each point. Standard error and confidence intervals were estimated by using 1000 iterations of Uno's perturbation-resampling technique. Intervals which contain 0 are denoted by shape. Although the models fit on the isoform-level filtering covariate set have a larger difference in concordance between isoform and gene-aggregate models, the standard error is greater. The points are slightly jittered to show overlap.

3.3 Final Models

We have two types of models we wish to compare and two sets of covariates they were fit on. We refer to these four categories by shorthand, where the first designation gives the pre-selection covariate group and the second gives the granularity of the model covariates. Iso-iso models are fit on the covariates pre-selected by isoform-level filtering and genomic covariates are represented by isoform log2 CPM ("iso" type models). Iso-gene models are fit on the set of gene-aggregated covariates ("gene" type models) pre-selected by isoform-level filtering. Gene-iso and gene-gene models are similar but fit on the set of covariates pre-selected by gene-level filtering.

We are particularly interested in iso-iso models and gene-gene models. Gene-gene models represent the pre-selection and model fitting possible before next-generation sequencing techniques, while iso-iso models represent the finer grained pre-selection and fitting possible with RSEM.

We found that fixing $\alpha = 0$ and searching over λ minimized the cross-validated loss for both of these model categories. The resulting estimated models are equivalent to those found using the ridge estimator. We focus our discussion of concordance on these ridge models for the iso-iso and gene-gene categories. We refer to these ridge models as the iso-iso and gene-gene models.

The c-statistic for the iso-iso model is higher than the gene-gene model. This suggests that models created and fit on the finer grained covariates have increased predictive power. We used the asymptotic normality of Uno's C estimator to construct a 95% confidence interval for the difference in concordance between the iso-iso and gene-gene models. The difference, presented in Table 15, was statistically significant. Again, this suggests that isoform counts provide increased predictive power compared to gene-level counts.

We present Kaplan-Meier estimated survivor curves per model for patients strat-

Model	Est	SE	Lower95	Upper95
Gene-Gene	0.89	0.02	0.85	0.93
Iso-Iso	0.94	0.01	0.92	0.96
Difference	-0.05	0.02	-0.09	-0.02

Table 15. Estimates of Uno's C for the iso-iso and gene-gene models, along with 95% confidence intervals. The 95% confidence interval for the difference in estimates, based on the asymptotic normality of the estimator, suggests the difference in estimates is statistically significant.

ified by median risk score in Figures 19 and 20. The curves are estimated using the same observations used to fit the models. As is the case with the concordance estimates, the curves likely overestimate model performance. However, we assume the difference between the curves is still indicative of the true difference between models. Test statistics from log-rank tests on the stratifications are given in Table 16.

Model	$\chi^2_{df=1}$
Iso-Iso	198
Gene-Gene	142

Table 16. Test statistics for log-rank tests for the final ridge models. Tests were on the association between survival outcome and whether the patient's risk score was below the median risk but used the same set of observations the models were fit on. However, there is a large difference in the χ^2 values between the two models.

3.4 Discussion

We present a short summary of the work performed and discuss the final models. Shortcomings in our methodology and further work are outlined.

3.4.1 Summary

Our intention is to investigate whether isoform-level expression information improves cancer survival predictions as compared to overall gene expression.

Our investigation is focused on women with breast cancer from the TCGA's BRCA datasets. One dataset contained high level clinical covariates and survival time for 1097 patients. Another contained RSEM estimated isoform counts sequenced from

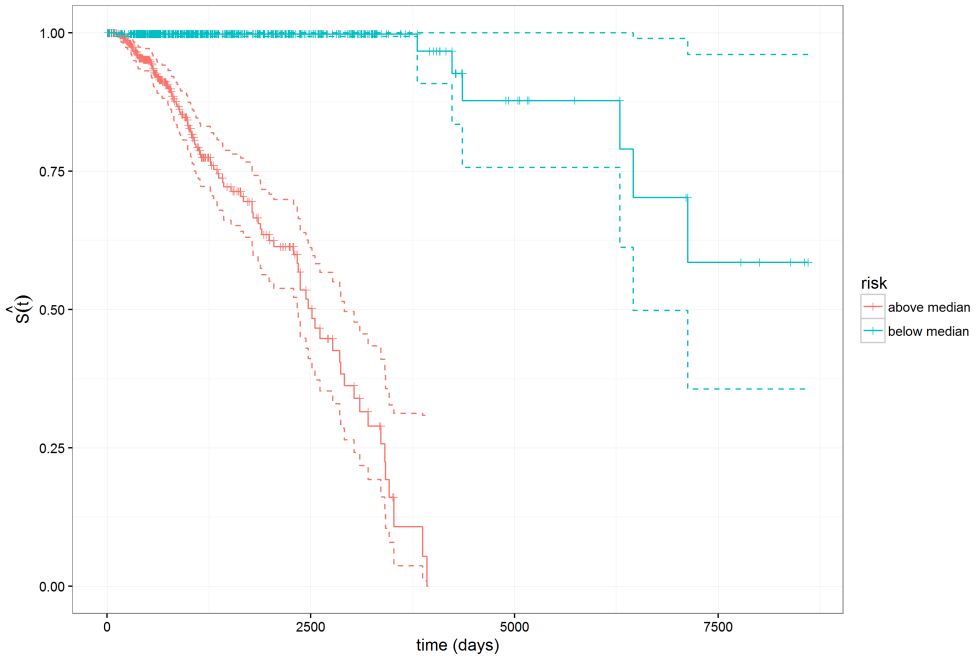

Figure 19. Kaplan-Meier estimate of the survivor functions $S(\hat{t})$ of patients stratified by risk score in the iso-iso ridge model.

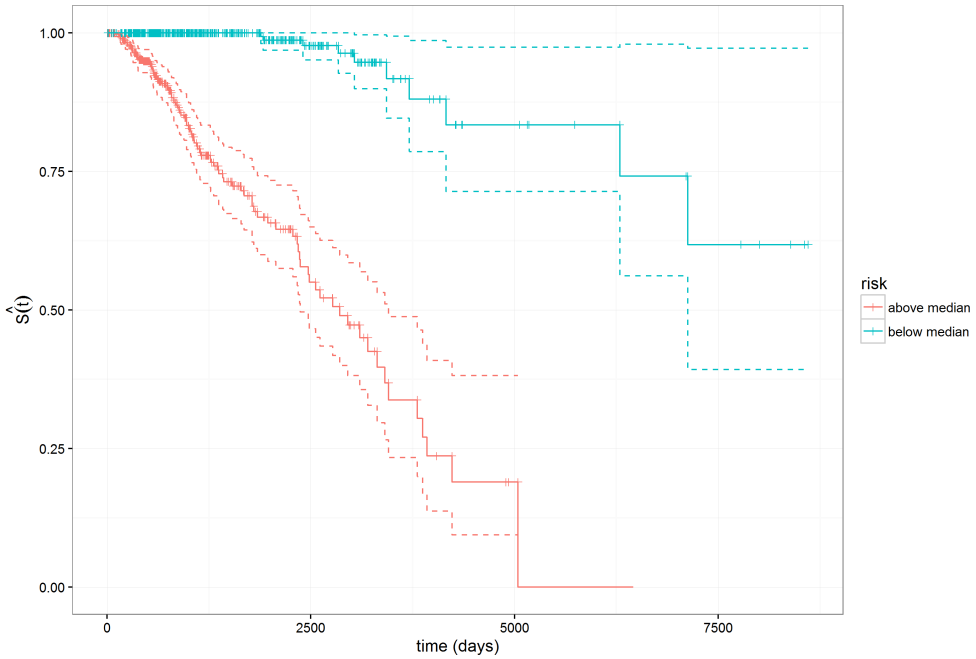

Figure 20. Kaplan-Meier estimate of the survivor functions $S(\hat{t})$ of patients stratified by risk score in the gene-gene ridge model.

1212 tissue samples, of which 1100 belong to breast tumors. We found that patient age, cancer stage, and the presence of radiation therapy were clinical covariates strongly associated with survival outcome.

The original RSEM data contained estimated counts for 73,559 isoforms per sample. However, many of these estimated counts were always 0 or exhibited extremely low variance. We took the sample variance of each isoform and ignored isoforms with a sample variance less than or equal to 1.

We normalized the remaining RSEM data between cancerous tissue samples by estimating and correcting for the "sequencing depth". We then added one to all corrected counts and converted them to CPM, or counts per million. Isoforms without a known gene correspondence were ignored, leaving 63,213 isoform (or 19,330 gene-level aggregate) counts for each sample.

Merging the two datasets yielded 967 observations of women breast cancer patients. 109 women died under observation. We performed variable pre-selection at the isoform and the gene-aggregate levels yielding two sets of covariates ("isoform-level filtering" and "gene-level filtering"). Pre-selection involved fitting a univariate Cox model for each covariate and collecting the p-value associated with a likelihood ratio test of the model. The set of p-values were corrected to control for the false discovery rate, and we selected the covariates which had a corrected p-value below 0.20.

Our models are fit on either the log2 transform of CPM ("iso" type models) or the log2 transform of CPM aggregates ("gene" type models), along with the clinical covariates. We performed a grid search to minimize cross-validated loss over the free parameters, α and λ, of the elastic net penalized Cox model. When $\alpha = 0$ the penalty term reduces to the ridge penalty, and when $\alpha = 1$ the model fits the LASSO.

We found that the ridge models minimized the cross-validated loss for our model

categories of interest, “iso” models fit to covariates determined by “isoform-level filtering” and “gene” models fit to covariates determined “gene-level filtering”. We computed estimates of concordance for the ridge models that minimized cross-validated loss and tested the difference in concordance for significance. We also estimated Kaplan-Meier survival curves for each model, stratifying observations by relationship to the median risk score.

3.4.2 Investigation of Final Models

Our results suggest there is a difference in the predictive power between our final models. The concordance estimate for the iso-iso model is higher than for the gene-gene model. When we created a 95% confidence interval for the difference using the asymptotic normality of Uno’s C, we found the interval did not contain 0 and that the difference was statistically significant.

Ridge models do not perform variable selection and estimate non-zero effects for all covariates. This suggests the difference in concordances is partially due to the difference between the univariate filtering procedures. The filtering procedures selected two sets of genes with non-trivial differences. 43 common genes were selected by both procedures. This leaves 33 different genes out of the 76 selected by only the isoform-level filtering, and 255 different genes out of the 298 selected by only the gene-level filtering procedure.

It is of particular interest that the iso-iso model had a higher concordance estimate than the gene-gene model. The isoform filtering showed a higher proportion of “unimportant” isoforms. Additionally, isoform counts exhibit greater variability than genes, since the latter are aggregate counts. The results suggest that despite this increase in noise and variability, isoform level information may provide meaningful advantages over gene-level information.

Furthermore, the concordance estimates suggest that both of these models

have increased predictive performance compared to models fit on clinical covariates alone. Repeating our methodology on just the clinical covariates yielded a ridge model which minimized the cross-validated loss across the grid search of α and λ. The estimated c-statistic of this model was 0.79 and the value of Uno's C was 0.717, with a 95% confidence interval of (0.649,0.785). Both final models with genomic covariates had statistically significant higher estimates of concordance, validating our assumption that genomic information has a non-trivial association with breast cancer survival.

3.4.3 Shortcomings

We used the same observations to fit each model and then to compute the concordance measurements. An ideal methodology would have computed the concordance on a separate set of observations. However, we would have to create this set by partitioning our data and this would have caused some issues due to the low number of events within the dataset. Our dataset had 109 observations with a known failure time. Reducing this number would introduce non-trivial additional variation in the estimated models dependent on the choice of partitioning. Similarly, concordance is estimated using the relationship between observations with at least one known failure time. Again, the choice of partitioning would introduce non-trivial variability into the estimations of concordance. We could account for this variability by systematically estimating models and concordances for all possible partitions but were unable to perform this due to computing constraints.

We rationalize our results due to our focus on model comparisons. The use of the same observations to measure model performance means we overestimated the performance. However, we aren't interested in the actual model performance but the differences between models. We compared models using the predicted risk scores of each model on the same set of observations as would be done under a partitioning

scheme. We assume that though each model's performance is overestimated, the relative difference is still a valid estimate of the true difference in predictive power.

Another important caveat with this work stems from the assumption that the underlying RSEM count values are uncorrelated between observations and are free of systematic error. Our data (and other available TCGA datasets) are the result of rolling collection by a number of participating centers. Recent work has shown that inadvertent biases introduced per center or batch of samples are not accounted for by RSEM and can introduce false positives in work that relies on RSEM counts (Love et al., 2015).

3.4.4 Future Work

Much of the work done is related to the preprocessing, cleaning, and filtering of the data. A thorough investigation into the additional predictive power of isoform-level covariates on cancer survival should include additional TCGA datasets following the same procedure. A comparison of our results with a replication on non-TCGA breast cancer survival data could also yield interesting findings.

Alternatively, testing the discriminative power of concordance using a known data-generating process would aid in the interpretation of results on real world data. This could also be used to validate our preprocessing and filtering techniques.

List of References

Love, M. I., Hogenesch, J. B., and Irizarry, R. A. (2015). Modeling of rna-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *bioRxiv*, page 025767.

BIBLIOGRAPHY

Bishop, J. M., "The molecular genetics of cancer," *Science*, vol. 235, no. 4786, pp. 305–311, 1987.

Broad Institute TCGA Genome Data Analysis Center, "Analysis-ready standardized tcga data from broad gdac firehose 2016_01_28 run," 2016. [Online]. Available: http://dx.doi.org/10.7908/C11G0KM9

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S., "Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments," *BMC bioinformatics*, vol. 11, no. 1, p. 1, 2010.

Cox, D. R., "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220, 1972.

Crick, F. *et al.*, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L., "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

Efron, B. and Tibshirani, R., "Empirical bayes methods and false discovery rates for microarrays," *Genetic epidemiology*, vol. 23, no. 1, pp. 70–86, 2002.

Fisher, L. D. and Lin, D. Y., "Time-dependent covariates in the cox proportional-hazards regression model," *Annual review of public health*, vol. 20, no. 1, pp. 145–157, 1999.

Friedman, J., Hastie, T., and Tibshirani, R., "glmnet: Lasso and elastic-net regularized generalized linear models. version1," 2013.

Gilbert, W., "Why genes in pieces?" *Nature*, vol. 271, no. 5645, p. 501, 1978.

Hanley, J. A. and McNeil, B. J., "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

Hastie, T., Tibshirani, R., and Wainwright, M., *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

Hoerl, A. E. and Kennard, R. W., "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

James, G., Witten, D., Hastie, T., and Tibshirani, R., *An introduction to statistical learning*. Springer, 2013, vol. 6.

Kleinbaum, D. G. and Klein, M., *Survival analysis*. Springer, 1996.

Koziol, J. A. and Jia, Z., "The concordance index c and the mann–whitney parameter pr (x> y) with randomly censored data," *Biometrical Journal*, vol. 51, no. 3, pp. 467–474, 2009.

Le Cessie, S. and Van Houwelingen, J. C., "Ridge estimators in logistic regression," *Applied statistics*, pp. 191–201, 1992.

Lee, K. L. and Mark, D. B., "Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, pp. 361–387, 1996.

Li, B. and Dewey, C. N., "Rsem: accurate transcript quantification from rna-seq data with or without a reference genome," *BMC bioinformatics*, vol. 12, no. 1, p. 1, 2011.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R., "Normalization, testing, and false discovery rate estimation for rna-sequencing data," *Biostatistics*, p. kxr031, 2011.

Love, M. I., Hogenesch, J. B., and Irizarry, R. A., "Modeling of rna-seq fragment sequence bias reduces systematic errors in transcript abundance estimation," *bioRxiv*, p. 025767, 2015.

Nelder, J. A. and Baker, R. J., "Generalized linear models," *Encyclopedia of Statistical Sciences*, 1972.

Pencina, M. J., D'Agostino, R. B., and Vasan, R. S., "Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond," *Statistics in medicine*, vol. 27, no. 2, pp. 157–172, 2008.

Pimentel, H. "In rna-seq, 2 != 2: Between-sample normalization." Accessed: 08/20/2016. 2014. [Online]. Available: https://haroldpimentel.wordpress.com/2014/12/08/in-rna-seq-2-2-between-sample-normalization/

R Development Core Team, "R: A language and environment for statistical computing," Vienna, Austria, version 3.2.5. [Online]. Available: http://www.r-project.org

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R., "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of statistical software*, vol. 39, no. 5, p. 1, 2011.

Strimmer, K., "fdrtool: a versatile r package for estimating local and tail area-based false discovery rates," *Bioinformatics*, vol. 24, no. 12, pp. 1461–1462, 2008.

Therneau, T., "A package for survival analysis in s." 2016.

Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tibshirani, R. *et al.*, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

Uno, H., "Package 'survc1'," 2013.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L., "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

Uno, H., Tian, L., Cai, T., Kohane, I. S., and Wei, L., "A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data," *Statistics in medicine*, vol. 32, no. 14, pp. 2430–2442, 2013.

Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L., "Survival prediction using gene expression data: a review and comparison," *Computational statistics & data analysis*, vol. 53, no. 5, pp. 1590–1603, 2009.

VanRossum, G. and Drake, F. L., *The Python Language Reference*. Python software foundation Amsterdam, Netherlands, 2010.

Verweij, P. J. and Van Houwelingen, H. C., "Penalized likelihood in cox regression," *Statistics in medicine*, vol. 13, no. 23-24, pp. 2427–2436, 1994.

Wilks, S. S., "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

Zou, H. and Hastie, T., "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.